An exploration of the language within Ofsted reports and their influence on primary school performance in mathematics: a mixed methods critical discourse analysis

Lawton, R

February 2023

Doctor of Education (EdD) Degree Thesis

An exploration of the language within Ofsted reports and their influence on primary school performance in mathematics: a mixed methods critical discourse analysis

Thesis submitted in accordance with the requirements of Liverpool Hope University for the degree of Doctor of Education, Rebecca Lawton, February 2023.

Table of Contents

Table of Figures

Appendices

Acknowledgements

With thanks to all those at Liverpool Hope University for their support and guidance, particularly my supervisory team, Sandra Hiett, Naomi Hodgson, and Cathal O’Siochru for their wisdom and encouragement.

With thanks to my family: Dave, Jack, and especially Charlie, without whom I could not have devoted the time and energy to complete the writing, for their unending support and patience.

Additional thanks to all those within school and around schools who contributed to the discourse, it was my conversations with you all that shaped the questions, the possible meanings and supported the analysis and evaluation. I hope the learning is of some use to us all going forward.

Abstract

An exploration of the language within Ofsted reports and its impact on primary school performance in mathematics: a mixed methods critical discourse analysis, Lawton, R.

This thesis contributes to the understanding of the language of Ofsted reports, their similarity to one another and associations between different terms used within 'areas for improvement' sections and subsequent outcomes for pupils. The research responds to concerns from serving headteachers that Ofsted reports are overly similar, do not capture the unique story of their school, and are unhelpful for improvement. In seeking to answer 'how similar are Ofsted reports' the study uses two tools, a plagiarism detection software (Turnitin) and a discourse analysis tool (NVivo) to identify trends within and across a large corpus of reports. The approach is based on critical discourse analysis (Van Dijk, 2009; Fairclough, 1989) but shaped in the form of practitioner enquiry seeking power in the form of impact on pupils and practitioners, rather than a more traditional, sociological application of the method.

The research found that in 2017, primary school section 5 Ofsted reports had more than half of their content exactly duplicated within other primary school inspection reports published that same year. Discourse analysis showed the quality assurance process overrode variables such as inspector designation, gender, or team size, leading to three distinct patterns of duplication: block duplication, self-referencing, and template writing. The most unique part of a report was found to be the 'area for improvement' section, which was tracked to externally verified outcomes for pupils using terms linked to 'mathematics'. Those required to improve mathematics in their areas for improvement improved progress and attainment in mathematics significantly more than national rates. These findings indicate that there was a positive correlation between the inspection reporting process and a beneficial impact on pupil outcomes in mathematics, and that the significant similarity of one report to another had no bearing on the usefulness of the report for school improvement purposes within this corpus.

Chapter One

Rationale

This thesis investigates the language and terminology of primary school Ofsted reports. These reports are the outcome of school inspection visits, of varying lengths, undertaken by the Office for Standards in Education (Ofsted) as part of the quality assurance process for education and childcare providers in England. The impetus for the study arose from my professional practice as a School Improvement Officer working with primary schools in England. In this role I found school leaders often read Ofsted reports as contentious. For example, school leaders, who rely on the areas for improvement section of reports in particular to inform their school improvement activity, would complain that these were excessively similar from one school to another. In my experience of working across many schools, a number of common complaints about reports emerged, including a perceived similarity of reports to one another, a perceived disparity of treatment when, despite the similarity, different judgements were reached, and a disparity between the wording of areas for improvement and the conversations that leaders remember having with inspectors in school. The focus of this research is to identify whether there are any trends in the language used to describe and evaluate schools, or in the language used to direct schools to improve.

In my role as a School Improvement Officer, I found school leaders tended to focus, understandably, on local reports, i.e., those of their own school and their main competitors or collaborators. This provides relevant details but gives them a limited view of reports in general, both across the country and over time. This study aims to clarify to what extent leaders' local perceptions are borne out at scale. Using a range of tools to investigate similarity and word frequency and to undertake sentiment analysis, the language within a large corpus of reports is investigated. Reports are coded into identifiable groups so that these groups may be compared to the whole corpus or to one another. The study considers the implications for individual schools, and for school improvement practice more generally, if leaders' concerns over similarity are supported by the large-scale analysis.

The study employs a method of critical discourse analysis that links the use of 'Big Data' methods, namely the ability to search thousands of sentences for a specific term using algorithms, alongside more qualitative, critical, and professional interrogation of the terms, both in context and as stand-alone references (Syed, 2020; Morris, 1971). This is to ensure that the study considers not only the concrete data on frequency or similarity, but also the use of terms in context, i.e., of the overall evaluation of or advice to a school (Lester et al, 2016). The impact of the inspection report is considered by analysing whether any terms and their use or repetition have any correlation with outcomes in schools following inspection. The language used in these reports is predicted to be critical as, in my role as a School Improvement Officer, school leaders described reports as having an influence on their decisions about which areas to prioritise for improvement before re-inspection, and there are mandatory, regulatory requirements for responses built into specific parts of the report. If there are particular terms or phrasing that lead to a higher grading on re-inspection, or that have a stronger influence on outcomes for pupils, it is imperative that school leaders and inspectors know this. Such findings could also change the practice of School Improvement Officers, whose role often entails the interpretation of reports to inform schools' action plans and targets.

The corpus of Ofsted reports included in this study includes Ofsted inspection reports on primary schools in England written between 1 January and 31 December 2017. This period represents the term of office of a single Chief Inspector. This selection both locates the corpus in a particular period, which can then be linked to associated data on school performance, and reflects a period not affected by significant influential external factors, such as radical government policy change or amendments to the inspection handbook or inspection practices. The detail of the practice of Ofsted inspection and inspectors in England is explored in Chapter 2.

Care has been taken to use only information that is already in the public domain, and to only include information that is released mandatorily as part of published reports, government publications, and performance data. Where possible, schools are described as averages and groups, e.g., 'schools graded inadequate'. No individuals are identified and where school names are used, it is in reference to the information on their report, which is publicly available and must be published on the school's own website as per government requirements at the current time. These restrictions form part of the ethical considerations undertaken in the design of the project, in order to comply with both university ethics guidelines and the BERA Ethical Guidelines for Educational Research (BERA, 2018) and to bring the project in line with similar studies. (Watson, 2001)

1.1 Ofsted, a brief history, and explanation of practice.

A brief history of Her Majesty's Inspectors (HMI) (1992-2018)

From the inception of schooling in the early eighteen hundreds, the government initiated Her Majesty's Inspectors of schools (HMI), initially to ensure the efficient spending of government money (Parr, 2020). Although at the inception largely from privileged, well-educated backgrounds, their increasingly liberal reports were designed to advise ministers on the state of education in publicly funded schools. In 1992, when there were fewer than 500 HMI, the Office for Standards in Education (Ofsted) was set up to regulate the largely unregulated activities of HMI that had been critical of government spending and required standardised reports and grading systems.

Historically, HMI were appointed by Her Majesty the Queen through an order of the privy council and formally distant to the Education Department of the time. This independence is still explicitly referenced in Ofsted's documentation and the Education Act (2005). In practice, the reports coming from 'external' inspectors (non-HMI) receive careful quality assurance from internal administration at Ofsted to ensure standardisation given the more distributed and fractional staffing structure. HMI reports, however, are normally published with minimal intervention, as a marker of the level of independence from the Ministry, '…No administrator would dare challenge an HMI's report'" (Kogan, 1971, p. 15). It is accepted that any inspector 'without fear or favour' can write and publish reports that are bound by policy but without bias (Allen, 1960, p. 236). This was originally secured through the HMI appointment process, accepting and highly training those with 'prestige, high reputation and expertise' (Kogan, 1971, p. 20).

These HMI were traditionally recruited from leadership positions in education across all phases. Over time, confidence in this approach has been eroded by the high turnover of fractional inspectors and perceived inconsistencies in inspector quality (Mogra, 2016). Published reports increasingly served as a proxy for the historical operational significance of the HMI, who had formerly linked practices and institutions by the nature of their role and influence. The reports and conversations held by inspectors increasingly fell outside of the control of the 'partnership between central and local government and the teaching profession' (Weaver, 1979) that had been a critical aspect of the HMI role. During the 1970s the HMI role had included providing advice and guidance outside of established policy directions, which led to the government questioning their autonomy. In the 1980s, the then minister for education, Kenneth Baker, challenged HMI for being 'anti-excellence, antiselection and anti-market' (Baker, 1993) due to their support for humanist, progressive practices.

During the late 1980s and 1990s, the government grew dissatisfied with HMI reports that were implicitly critical of government policies, and public think tanks chided inspection procedures for being insufficient and for subscribing to 'trendy' educational theories (Maclure, 1998). The problem was described as an intelligence deficit at policy level, with policy makers lacking a concrete evidence base, although by 1989 HMI were publicly involved in policy making which led to serious questions over the independence of inspection. The replacement of HMI by the Office for Standards in Education ('Ofsted') in 1992 was an attempt to re-separate the inspectorate from the government. The aim was to reduce the influence of HMI on policy, standardise inspection activities and reports, and clarify the role of inspectors as 'referee' rather than 'coach' in a new government architecture that challenged liberal or permissive approaches, and which now depended heavily on routinised scrutiny, evaluation, and audit (Lee and Fitz, 1998).

1.2 Ofsted

Across Europe, Ofsted is one of the most 'mature' inspection regimes. Its role as an enforcer of standards across a distributed system, with multiple, 'independent' inspectors operating within a given framework, has been replicated in many Western countries, along with the adoption of the New Public Management rationality of baseline measures and standardised performance indicators to drive permanent improvement, illustrated by the Swedish schools Inspectorate (2010) belief that: "It has to be possible for all pupils to attain all objectives – to 100 percent" (p17). (See chapter 2.1: comparable inspection structures).

For this style of system to function, local, simplified, static and centrally controlled knowledge (such as test results, or attendance percentages) are shared with the public and nationally standardised. 'Post-bureaucratic' governance, described as a more advanced approach, instead involves de-centralised, autonomous, fluid, co-produced knowledge from networks of policy makers, experts, and practitioners, leading to an easy exchange and marketisation of education within neo-liberal economies (Thrift, 2005). Here, innovation is valued over consolidation and policy is formed by practice rather than policy informing practice. These 'quality-regimes' (Dahler-Larsen, 2011) although described as advanced compared to current systems, are yet to be found in common practice across Europe, as governments are reluctant to relinquish systems with central power and control.

Ofsted is the current mechanism used to inform the Secretary of State for Education of the quality of education in schools in England (and Wales). Its structure and functions are a response to the Education Act (2005), which made it mandatory for all maintained schools in England to be inspected. Prior to the Education and Inspections Act (2006), the function of Ofsted was to report on standards, leadership and management, behaviour, and attendance. The requirement for an annual report by the Chief Inspector and 'additional reports where needed' became law. The 2006 Act gave Ofsted a 'right of entry' into schools, including access to records and pupils and set a regulation for a time limit between inspections. It also added Sections 16 to 119 to the previous Act, which included the 'encouraging' of improvement.

> (1) The Office is to perform its functions for the general purpose of encouraging—
> (a) the improvement of activities within the Chief Inspector's remit
>
> Section 117 (1)(a), Education and Inspection Act (2006)

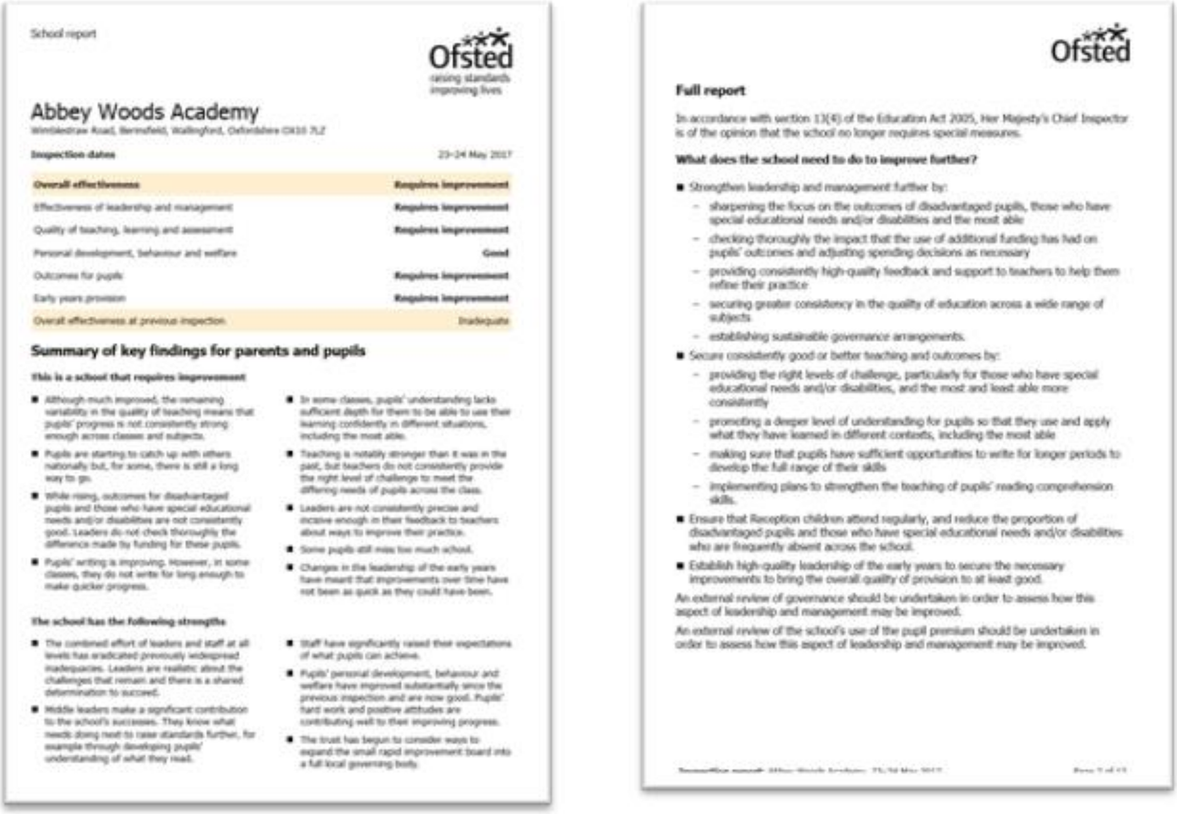
Previously, their remit only covered continued monitoring of those schools in categories of concern and overseeing the potential closure in the event of no improvement being evident. Over time, changes have been made to the Core Framework and Ofsted's activities and publications in response to amendments to the Act or the wishes of different Chief Inspectors (HMCI).

Currently, the inspection process involves a school visit, which culminates in a published, written report and an overall judgement of either Grade 1 (outstanding), Grade 2 (good), Grade 3 (requires improvement), or Grade 4 (inadequate), based on assessment of six core areas: effectiveness of leadership and management; quality of teaching, learning and assessment; personal development, behaviour and welfare; outcomes for pupils; early years provision; and overall effectiveness. As witnessed in my professional practice, and in line with the Education Act (2005), inspection visits and reports vary in length and purpose, depending on the previous evaluation of the school. Since 2015, schools judged to be providing at least a 'good' education were subject to shorter, single day visits, and letters rather than reports (HMCI, 2015), under the additional Section 8 of the Act (2005). Those inspections undertaken within the standard time frame associated with their previous judgement, or those triggered by safeguarding concerns in line with the requirements of the original Act, result in a full report under Section 5. This research focuses on full 'Section 5' reports on primary schools in 2017, which have a common structure.

In November 2017, there was a formal amendment to the basis on which inspections were determined to be Section 5 or Section 8. In addition to safeguarding risks and normal timescales, those ‘good’ schools who were at risk of not retaining their previous grade due to a notable change in outcomes or profile, would undergo a Section 5 inspection rather than the shorter Section 8. This had the effect of increasing the proportion of Section 5 inspections undertaken. A March 2018 Ofsted Statistical Release stated that “96% of inspections of good schools in September and October 2017 were short inspections, compared to 75% between November 2017 and April 2018” (Ofsted, March 2018 Statistical release).

Section 5 reports include a front-page judgement of each of six core areas listed (see Figure 1), with a summary bullet points list of key strengths or weaknesses, depending on the overall effectiveness of the school. On the second page is a list of recommendations for improvement, and the main body of the report covers each of the six areas in detail, outlining performance against national expectations, comparing the school’s performance to the national picture, and highlighting areas to improve. Inspectors have a limited word count, and reports are generally of a similar length, although where a school is found to be inadequate reports can be much longer and can take months to reach publication, whereas other reports are normally published within a month of the on-site visit.

The literature around Ofsted and the use of HMI describes an acceptance that, although labelled as ‘independent’, HMI are used as the ‘voice of the inspectorate’ and regularly steer national policy (Wood, 2019). Ofsted reports have been explicitly designed to be a tool for improvement, and subsequent inspection processes have been amended to enable this over time, with quality assurance and bureaucratic processes standardising practice around reported elements, judgements, and improvement guidance (Ofsted Handbook, 2017).



School report

Ofsted
raising standards
improving lives

Abbey Woods Academy

Wimblestraw Road, Berinsfield, Wallingford, Oxfordshire OX10 7LZ

Inspection dates	23–24 May 2017
Overall effectiveness	**Requires improvement**
Effectiveness of leadership and management	**Requires improvement**
Quality of teaching, learning and assessment	**Requires improvement**
Personal development, behaviour and welfare	**Good**
Outcomes for pupils	**Requires improvement**
Early years provision	**Requires improvement**
Overall effectiveness at previous inspection	Inadequate

Summary of key findings for parents and pupils

This is a school that requires improvement

- Although much improved, the remaining variability in the quality of teaching means that pupils' progress is not consistently strong enough across classes and subjects.
- Pupils are starting to catch up with others nationally but, for some, there is still a long way to go.
- While rising, outcomes for disadvantaged pupils and those who have special educational needs and/or disabilities are not consistently good. Leaders do not check thoroughly the difference made by funding for these pupils.
- Pupils' writing is improving. However, in some classes, they do not write for long enough to make quicker progress.
- In some classes, pupils' understanding lacks sufficient depth for them to be able to use their learning confidently in different situations, including the most able.
- Teaching is notably stronger than it was in the past, but teachers do not consistently provide the right level of challenge to meet the differing needs of pupils across the class.
- Leaders are not consistently precise and incisive enough in their feedback to teachers about ways to improve their practice.
- Some pupils still miss too much school.
- Changes in the leadership of the early years have meant that improvements over time have not been as quick as they could have been.

The school has the following strengths

- The combined effort of leaders and staff at all levels has eradicated previously widespread inadequacies. Leaders are realistic about the challenges that remain and there is a shared determination to succeed.
- Middle leaders make a significant contribution to the school's successes. They know what needs doing next to raise standards further, for example through developing pupils' understanding of what they read.
- Staff have significantly raised their expectations of what pupils can achieve.
- Pupils' personal development, behaviour and welfare have improved substantially since the previous inspection and are now good. Pupils' hard work and positive attitudes are contributing well to their improving progress.
- The trust has begun to consider ways to expand the small rapid improvement board into a full local governing body.

Ofsted

Full report

In accordance with section 13(4) of the Education Act 2005, Her Majesty's Chief Inspector is of the opinion that the school no longer requires special measures.

What does the school need to do to improve further?

- Strengthen leadership and management further by:
  - sharpening the focus on the outcomes of disadvantaged pupils, those who have special educational needs and/or disabilities and the most able
  - checking thoroughly the impact that the use of additional funding has had on pupils' outcomes and adjusting spending decisions as necessary
  - providing consistently high-quality feedback and support to teachers to help them refine their practice
  - securing greater consistency in the quality of education across a wide range of subjects
  - establishing sustainable governance arrangements.
- Secure consistently good or better teaching and outcomes by:
  - providing the right levels of challenge, particularly for those who have special educational needs and/or disabilities, and the most and least able more consistently
  - promoting a deeper level of understanding for pupils so that they use and apply what they have learned in different contexts, including the most able
  - making sure that pupils have sufficient opportunities to write for longer periods to develop the full range of their skills
  - implementing plans to strengthen the teaching of pupils' reading comprehension skills.
- Ensure that Reception children attend regularly, and reduce the proportion of disadvantaged pupils and those who have special educational needs and/or disabilities who are frequently absent across the school.
- Establish high-quality leadership of the early years to secure the necessary improvements to bring the overall quality of provision to at least good.

An external review of governance should be undertaken in order to assess how this aspect of leadership and management may be improved.

An external review of the school's use of the pupil premium should be undertaken in order to assess how this aspect of leadership and management may be improved.

Inspection report: Abbey Woods Academy, 23–24 May 2017 Page [illegible]

Figure 1 Extract from a Section 5 Ofsted Inspection report, 2017

Ofsted's inspections of schools perform three essential functions. They:

- provide parents/carers with an expert and independent assessment of how well a school is performing, and help inform those who are choosing a school for their child
- provide information to the Secretary of State for Education and to Parliament about the work of schools and the extent to which an acceptable standard of education is being provided. This provides assurance that minimum standards are being met, provides confidence in the use of public money and assists accountability, as well as indicating where improvements are needed.
- promote the improvement of individual schools and the education system as a whole.

From "The Framework for School Inspection" Ofsted, (2013) p4

The inspection process outlined in the Ofsted handbook involves consultation and formation of the basic format of the final evaluation statements in partnership with school stakeholders. The last stages of an inspection visit are described as where the point at which inspectors and school leaders discuss and agree the outline content of the subsequent report is

discussed and agreed and confirm that the written report will only ‘differ slightly’ from this (Ofsted Handbook, 2017). The lead inspector then writes the report away from the school site, and it is processed through a series of quality assurance and proofing systems, for which the specific content and procedures themselves are not in the public domain.
Parts of the quality assurance system that are public knowledge include a draft reading of the report by school leaders and governors after Ofsted’s internal assurance process and prior to publication, so that leaders can make corrections or challenge request amendments to the agreed content. Breaking of confidentiality regarding the report prior to publication can result in a demotion of the leadership and management judgement and is considered a serious breach of professional conduct.

The structure of the thesis starts with Chapter 2 as a literature review, looking at existing inspection structures and their comparable methodologies and international correlations. Chapter 2 concludes by summarising five core research questions. Chapter 3 describes the methodology, including my positionality and the design, tools and assumptions within the research. The model and ethical considerations are included within this chapter. Chapter 4 covers the empirical findings, from each of the software analyses and some initial theories and refinements during the iterative process. Several groups of data are considered including report author variables, type of inspection and prior gradings of schools. One line of enquiry is selected for comparison to outcomes for pupils, and the findings of this, and the other enquiries discussed in Chapter 5.

Chapter Two

Literature Review

The literature review covers comparable school inspection structures and methodologies across Europe, findings from other studies looking at these structures and detail on the reporting of inspection, the English system and associated quality assurance methods within education

2.1 Comparable Inspection structures

Inspection practices in education have been in place since the early 19th century in the Irish, Czech, Austrian, Swedish, and Dutch educational systems (Greger, 2011; Schiepl and Seel, 1985; Lindgren 2014). The design and implementation of these over time have all included reference to educational effectiveness and quality plus inclusion of legal and administrative appropriateness. The pressures and opportunities of globalisation and the systems of comparison and competition that it has generated internationally have led to inspection systems referred to in terms of 'evidence-based governance' (Altrichter & Kemethofer, 2015) where monitoring and control is enacted through the collecting and analysing of performance audits. These 'evidence-based' systems appear to be built on assumptions about intermediary mechanisms and processes, embedding audits, evaluation, and accountability but with culturally specific variations (Eurydice 2004; Maag Merki, 2010). These evidence-led systems of inspection have similar structures across Europe, setting expectations by describing standards and procedures, collecting evidence via visits and published data, and holding schools accountable for student achievement, teaching, organisation, and leadership. The stimulation of school and system improvement is realised by publishing reports highlighting strengths and weaknesses and giving recommendations that reflect national educational policy (Creemers et al, 2007).

As this project is so closely linked to the English lexicon, I have limited the scope of my discourse analysis method to those aspects relevant to the English language. However, there are some notable comparable international reports and projects that inform the context, interpretation, and methodology. Several studies have explored the impact and purpose of school inspection in Europe (including the UK), which follow a similar inspection structure to that described above (Coffield, 2009; Ehren, 2014; Janssens and Maassen, 2014; Janssens and van Amelsvoort 2008; Lindgren et al 2012). Some studies focus on Ofsted specifically (Jones and Tymms, 2014; Richards, 2012 and 2016; Watson, 2001; Ozga and Lawn, 2014). Ehren (2014) found that inspections of failing schools in England (which are targeted for more frequent inspections in the Ofsted model) have the most impact, but the highest proportion of unintended consequences, such as teacher workload. Ehren found that although improved teaching conditions and capacity sometimes ensue, there was little evidence of this leading to improved outcomes for students. This meta-analysis also describes a clear difference between the impact of inspection on primary and secondary settings. Bellmann and Weiß (2009) identify more than 20 unintended effects of inspection, including a loss of trust, narrowing of the curriculum, discouraging new teaching strategies, damage to work satisfaction, and cheating by students and teachers (e.g., Kotthoff, 2003; Maag Merki, 2010; Perryman, 2007). It has also been stated that less effective schools "did not manage to improve their status simply because of the pressure placed upon them" (Creemers & Kyriakides, 2012, p. 51; also, Good, Wiley, & Sabers, 2010).

In England, the inspection process is open to public and academic scrutiny (Richards, 2012) and has been criticised for, among other things, its negative impact on teachers (Case, Case and Catling, 2000), its impact on headteachers (Courtney, 2013), of being a political tool that embeds competition rather than equity (Lefstein, 2013), and of driving leaders in underperforming schools to attempt to 'game the system' (Ouston, 1997; TES, 2017). These, often anecdotal, case study or small-scale findings are generally borne out to varying extents by the Eurydice meta-analysis (Ehrens, 2014), which looks at a wide range of quality assurance practices in education across Europe. However, the lingering criticism that there are no established overarching techniques to measure the impact of inspection on improvement (Fitz-Gibbon, 1999; Earley, 1998) appear to be valid.

Jones and Tymms (2014) developed a conceptual model that attempts to describe the assumptions underpinning the English inspection regime and its impact on school improvement, in a way that mirrors the Eurydice study. Their claims to validity are supported by the inclusion of senior Ofsted officials directly within the evidence gathering and conclusions, which implies a degree of 'insider' ratification. These pieces of research, amongst a myriad of others citing inspection as a factor when focussing on other aspects of schools, describe the heavy weighting of the emotional and social aspects of the inspection process, and the important role of the inspector within the structure.

Current inspection systems across Europe focus on the evaluative and administrative functions of leadership and 'professionalise' the process by formalising the role of the inspector, emphasising the need for 'expertise' and using academic and scientific terminologies and instruments. A number of literature reviews have attempted to source meta-findings about the impact of modern inspection processes on improvement (De Wolf &

Janssens, 2007; Ehrens, 2014; Ehren & Visscher, 2008; Husfeldt, 2011; Klerks, 2013; Kotthoff & Böttcher, 2010; Luginbuhl, Webbink, & De Wolf, 2009).

Husfeldt (2011, p. 260) describes three types of study commonly undertaken:

> '(1) descriptive studies about attitudes and expectations with regard to inspections' describing relationships, pressure, and impact on staff.
>
> '(2) Descriptive studies about reactions to inspections'
> commonly using qualitative or quantitative methods to analyse actions undertaken following inspection, such as improvement programs, and their consistency or proportionate scale.
>
> '(3) Studies about improvement of student performance after inspections'
> Normally linked to concrete measures such as exam performance.

These meta-analyses indicate that changes in performance following inspection are generally small and can be positive or negative, with variance impacted by outcome and prior grading.

Husfeldt argues that the reason research findings were inconclusive was due to a lack of a theoretical model that mediated between inspection and school processes, and which could cover the sum of influencing variables. Ehrens, in the 2014 report, proposed a conceptual framework for understanding inspection processes, contexts, and results and used this to compare six European inspectorates by using legal and administrative documents and interviews with officials. This process resulted in consensus around major goals and processes across countries (setting expectations, accepting feedback, actions of stakeholders), which indicated 'effectiveness mechanisms.'

Altrichter (2015) looked at variables such as 'resources for improvement' and 'educational goals', which had no marked impact following inspection, whereas 'pressure to improve', stemming from political and procedural forces as well as stakeholders and competition, did have a measurable impact. These were linked to public sanctions for underperformance in both case studies and meta-analyses (Chiang, 2009; Perryman, 2010; Kotthoff et al, 2007; Van Bruggen, 2010; Faubert, 2009).

This use of inspection as an instrument to deter schools from deviating from the regulatory authorities' 'legitimate' or 'acceptable' approaches is discussed by Braithwaite (2008) as a form of 'regulatory capitalism'. These inspection regimes include a focus on standards, including compliance with legal regulations (hours of schooling, equality, security of information), and processes (subject teaching, relationships), and most include results in statutory tests and examinations. Most have a ‘critical threshold’ or minimum qualifying standard and some form of quality measure. All are designed to be consequential, and to stimulate and orientate improvement.

Ehren and Visscher (2006) state that accountability is understood to drive improvement, in that actions will follow a judgement of underperformance against accepted standards. Research on the English inspection system does not reflect an improvement in the quality of teaching and learning following inspection (Earley, 1998; Gray and Wilcox, 1995; Kogan and Maden, 1999); a decline in standards has been reported in some cases (Shaw et al., 2003; Rosenthal, 2004), although some improvement in the weakest settings has been shown (Matthews and Sammons, 2004) and some ‘change’ to provision does generally occur (Wilcox and Gray, 1996). Wilcox describes this change as only substantial when leaders see inspection recommendations as a validation of existing plans and ideas. Several studies cite negative effects, such as increased stress and workload, a reluctance to change (Chapman, 2001; Gray and Gardner, 1999; Leeuw, 1995, 2000), and potential fraudulent manipulation of data (Wiebes, 1998).

Research focused on European systems of inspection, generally informed by evidence-based methods of school governance that link the use performance measures to national policy, shows that inspection has some impact on school performance, particularly for inadequate schools (Ehrens, 2014), but even large meta-analyses include limited data on the impact of inspection on educational outcomes. Research on inspection has tended to focus on the emotional, administrative, and unintended consequences of inspection. There is limited research that measures impact by linking the act of ‘inspection’ to performance against a minimum national compliance threshold, and there is no link made between the different internal mechanisms of inspection - reporting, grading, structure, etc. - and differences in outcomes beyond small-scale case studies.

2.2 Reporting inspection findings

In some countries, inspection reports are made public, whereas in others they are only made available to key stakeholders. The structure of the inspection and reporting system across Europe is cited as having an impact on actions and outcomes, specifically the differences in structure and number of recommendations made (Wilcox and Gray, 1984) and the balance of direct (instruction) and indirect pressures, including the publication of judgements. Reports are considered an ‘indirect intervention’ rather than a direct intervention, such as verbal feedback in a meeting (Ehren and Visscher, 2006). For example, in the Netherlands, schools are only required to self-evaluate against government-set criteria. The reporting process is cited as the core ‘deterrent’ or sanction in European schools, as it is associated with public awareness of failings, position in rankings, or comparative performance.

The ‘working methods’ of school inspectors are described as a factor that influences how schools react to, and the side effects of, inspection (Ehren and Visscher, 2006). Previous studies have reported on the direct, one-sided nature of feedback during inspection (London, 1995) and the expectation of feedback (whether directed by the process or not) during the inspection (Scholtes et al, 2002), as well as the comparison to standards during feedback and structures for this feedback (Black and William, 1998) as influential factors. These studies find that “clear and explicit reports are more successful in informing school improvement plans” (Matthews and Sammons, 2004, p.45), and that specific, constructive feedback is the most effective (Brimblecombe et al., 1996; Archer-Kath et al, 1994) as long as the recipient believes this is an accurate picture of performance from a credible source (Ilgen et al., 1979).

Ehrens’ (2015) European research indicated that ‘differentiated models’ of inspection, where both a visit and public report are used, were most effective in securing improvement, although this also led to a narrowing of the curriculum and lack of innovation. The Netherlands’ approach of publicly listing failing schools, and publication of reports by the Netherlands, England, Sweden, and Ireland were compared to the Austrian system, which requires parental meetings following the report, and to the Czech Republic, where only thematic surveys are published. This indicated that public reporting generated more acceptance of feedback, yet those with sanctions were less accepting.

Similarly, Ouston’s (1997) findings demonstrated that school inspections promote greater school improvement if the school report details the respect(s) in which the school has performed poorly. Some research, such as Watson (2001) and Leite (2014), has focused on the documentation used to support inspection processes, and identifies the language of reporting as an area for further research. However, few studies have investigated the impact of terminology or language within Ofsted reports specifically, beyond a focus on the language of ‘feedback’, such as in Schweinberger et al (2017) or Ball (1997). Schweinberger reflects on communication methods used in inspection in Switzerland, and the process of interpretation from verbal and written feedback into actions. Ball (1997) describes policy language within these published reports as ‘policy in action’, a form of ‘micro-disciplinary practice’ in which inspection practices reinforce the language of educational policy and influence in-school activity and are therefore seen as an example of government as ‘steering from a distance’.

> Public reporting had a strong positive effect on ‘actions of stakeholders’, and on ‘improving self-evaluations’, and via indirect effects it also exerted influence on ‘change in capacity building’ and on ‘change in school effectiveness’. These results show that the different school inspection models are associated with a differentiated pattern of influence on the mechanisms that generate impact of school inspections. Thus, whether there is public reporting or not influences stakeholders’ actions directly, and it also directly influences the schools’ self-evaluations. These factors in turn have effects on the principals’ improvement actions.
>
> (Ehren et al, 2015, p. 388)

Ehrens’ large scale Eurydice project describes intended and unintended consequences of inspection, and concludes that a two-year short-term impact tends to follow an inspection visit, with a dynamic, non-linear, longer-term impact observed in inspection systems that use “a differentiated, high-stakes approach, focused on outcomes of schools” (2014, p. 5): Schools that move from a positive assessment to a negative assessment “…become less open to inspection feedback; failing schools indicate little sensitivity of stakeholders to their inspection report as a driver for change” (p23).

The Dutch and English publication of inspection findings are described as intended to enable parents to contribute to school improvement through their choices. It is expected that if the best schools are the most popular, others will be motivated to improve, and that parents would use inspection findings and comparisons to pressure their schools to improve.
This assumption that public reporting will encourage parental pressure is not borne out by the research (Dronkers and Veenstra, 2001; Educational Council, 2001; Karsten and Visscher, 2001), however, which shows that parents are more interested in aspects of schools not reflected in inspection reports, such as reputation, environment, and entrance requirements. Ehren and Visscher (2008) surveyed primary school inspectors and found that over half “did not think of parents as the main audience for school reports, and only four per cent of the school inspectors thought that parents should have a role in school improvement” (p.205).

In the case of Ofsted, reports are described on its website as tools to be used by policymakers to judge effectiveness and to monitor and improve the quality of education. However, in the publicly available Ofsted Handbook (2018), the report is described in terms of use by and availability to parents, and the rhetoric of ‘informing parental choice’ is found throughout the Ofsted handbook, supported by links to the Parent View website, designed by Ofsted to gather parental feedback.

This positioning of the inspection report as a tool for parents is also supported by the requirement for schools to publish their inspection report on their school website and the public availability of a database of inspection reports (Ofsted Handbook 2017, #134) where school reports can be compared. In the English system, once a report is published it is permanently available in the public domain and it is difficult for any stakeholder to get the report removed, changed, or replaced. Only in very rare cases where a complaint is substantial and on-going can a report be amended or removed (Ofsted Handbook and Complaints Guidance, 2017). Although publication of some reports is delayed, the vast majority are published within 28 days of the inspection taking place, as can be seen in the dates of inspection compared to dates of publication on the Ofsted website and in the annual report relevant to the sample (HMCI Report, 2015).

This body of literature suggests that the indirect intervention of publicly reporting a schools’ performance has more impact on school improvement than direct communication with inspectors. It suggests that, across Europe, a limited number of recommendations and judgments and the agreement on the detail of their content is the most effective tool to secure improvement. Acceptance by leaders of the accuracy of the judgements and comparative ranking of a school is a critical element in improving performance, and the wording of feedback or reference to compliance is often used by inspectorates to steer schools towards alignment with national policy.

2.3 Quality assurance in education

Education is subject to an increasing and changing array of internal and external forces acting upon it. Both in the day-to-day practices of formal education settings and the broader contexts of serving and being accountable to society, it is subject to objectification, modification, and evaluation at every level (Creasey, 2018).

Biesta (2009) argues that education meets three functions: qualification, socialisation, and subjectification. He describes its intrinsic (humanist) values, and extrinsic utilitarian neoliberal purpose as accepted across Europe today, and terms such as ‘human capital’ (Hartog and Oosterbeek, 2007) are commonplace, despite increasing evidence that the ‘rate of return’ on education has been falling for the past 50 years (Blacker, 2013). Recent research (such as Means, 2017) suggests that technological and economic developments should be steering education to more humanist or self-development models as the increase of automation in working life renders other measures of ‘successes’ obsolete.

The pressure on schools to consistently improve outcomes, reinforced by the mainstream press (Wiggins, 2015 and Philips, 1999), amplifies the utilitarian approach to education quality. This reflects an increasing discourse of business and management in education, where descriptions such as ‘coasting schools’ and ‘continuous improvement’ imply that schools previously evaluated as successful in terms of culture and behaviours should be equally striving for continuous increases in quantitative, comparable measures. Where a ‘national average’ is provided, this separates schools into those above and below, with ‘below’ being synonymous with failure, despite contextual factors such as starting points.

> Sir Michael Wilshaw, Head of Ofsted condemned the fact that one in five pupils are leaving primary school without reaching the national average in English.
>
> (Curtis, 2012, in *The Guardian)*

Adherence to this utilitarian methodology via formal national testing arrangements can shift the focus of school curriculum from humanist principles, i.e. a focus on those things that are good for children to learn, to a focus on those things that are demonstrable within assessment. As institutions are evaluated by these measures and subsequently deemed to be successful or ‘failing’ on this basis, it is understandable that leaders structure school improvement plans around these criteria in particular. Ofsted reports have been criticised for over reliance on these ‘data’ in the past, and the retention of descriptive inspection commentary within reports reflects the attempt to balance values attributed to not only concrete measures, but more general principles when evaluating a school’s effectiveness. This also sustains the mantle of the inspector as expert, interpreting policy for school leaders, while upholding the consistency of the business model of inspection (to time, to budget, within word count).

The relationship between evaluation structures and governance has been a significant areas of research in recent years (Grek and Lindgren 2014; Ball, 2003; Segerholm and Åström, 2007), whereby the information gathered via inspection is seen to be used as regulatory performance measures that support the centralisation of power through a ‘performance-evaluation nexus’ (Clarke, 2004) where degrees of autonomy are earned by individual institutions through compliance with performance benchmarks (Lawn, 2006). This is evident in the Ofsted framework, for example, as inspection frequency is reduced for schools judged good or outstanding (Ofsted Handbook, 2017, p. 10).

Those schools unable to evidence good and improving outcomes for children in terms of assessment results are at risk of being judged as failing to provide a good education. Given what is at stake in being judged to be failing, e.g., a loss of autonomy and possibly compulsory academisation (Education and Inspection Act, 2016), it makes sense that leaders opt for a 'teaching to the test' approach and strict adherence to Ofsted criteria. In cases where institutions continue to 'fail', increased state control in the form of financial penalty, external imposed governance or even closure is used to focus local policy and strategy on increasing outcomes (Creasey, 2018). Risk-avoidant leadership approaches that can result from such a 'high stakes' regime include changing levels of inclusion or selection protocols (admissions policies) or restricting the range of qualifications offered and are approaches that have to be regulated alongside outcomes in the name of equity (Adams, 2016; Finn, 2015; Coffield and Williamson, 2011; Stobart, 2008). This standardisation of the evaluation model in order to provide 'fair' and consistent criteria can lead to standardisation of educational provision, thereby limiting the breadth and range of what is available for children, and potentially discriminating against some settings - for example, schools with high proportions of special needs pupils - who do not fare well in this competitive model of national, standardised measurement.

These moves towards quantifying education increases the power and influence of mediating external regulatory bodies such as Ofsted, exam boards and so on, the power of auditing, the influence of standardisation practices, and the impact of change at policy level (Hussey and Smith, 2010; Power, 2003). In this target-driven culture, leadership tends towards a risk-averse approach (Bloom, 2017; Biesta, 2013). Some phrases within the descriptive sections of the Ofsted handbook such as 'broad and balanced curriculum' and reporting specifically on provision for those children with special educational needs and those in Alternative Provision settings, could be seen as an attempt to balance quantitative with qualitative judgements (Ofsted Handbook, 2017, p. 47).

It is safer within this structure for inspectors to identify areas for improvement based on quantitative measures, for which there is an objective, agreed meaning within the system, and for school leaders to comply with these, rather than to offer more holistic, cultural or behavioural targets. This is further compounded by the increased use of serving headteachers as lead inspectors, who, conditioned by the risk-averse culture, perpetuate accepted norms. For example, 50 of the most prolific lead inspectors had never awarded 'outstanding' status during their time as school leaders (Exley, 2015), and the Policy Exchange think tank (Waldegrave, 2014) criticised Ofsted in 2014 for using data as a 'safety net' to standardise inspection judgement reliability. Research shows that despite changes in approach or framework, inspectors try to situate current inspection practice into a narrative of continuity.

> This attempt to hold on to, and make use of, the past - or the imagined past - of inspection shows the continuing strength of embedded practices and assumptions in the nation that are historically shaped and that have framed the assumptive worlds and practices of inspection.
>
> McPherson and Raab (1988) p58 'Governing by Inspection', Grek et al 2015)

Combining the personal or humanist with the objective or instrumental is referred in the handbook (Ofsted Handbook, 2017, p. 38, #136) as the use of 'professional judgement' and relies on inspectors being 'expert' in their field, and experts in translation of 'evidence' into 'knowledge':

> Inspectors bring their expert judgement and 'objective' data into relationship with one another, within more or less prescribed parameters; they are responsible for making knowledge about system performance available for translation into use by policy makers at all levels, and by practitioners; and they are also engaged in building knowledge about improvement within and across systems. At the same time, inspectors are responsible for ensuring that (sometimes shifting) accountability requirements are met to greater or lesser degrees; they claim independence from central governments and offer public judgements about the performances of education systems that have political implications.
>
> (Clarke, 2005, in Grek and Lindgren, 2014 p6.)

Here, the inspector is the conduit of power, turning the evidence of inspection, the snapshot impression of a single institution, into 'knowledge' about education provision in general, using a format accessible to both governments and schools. Their task is to remain independent of the government that employs them when they describe in their report the impact and success of the directives imposed by that government. During the timescale of this project, the use of sub-contractors to employ inspectors had been removed, and 'serving practitioners' from schools judged at least good were directly contracted to undertake inspections. This was a continuation of the reduction of directly employed HMI. When Ofsted was established in 1992, the number of HMI dropped from 515 to less than 300 and was further reduced in 2016. The recruitment of large numbers of additional 'Ofsted' inspectors on zero-hour contracts reflected the industrial scale of inspection, which by 2009 also included teacher training and independent settings, thereby vastly increasing the influence of the inspectorate, and Ofsted's criteria, in all areas of education.

This body of literature describes the inspection process in England as becoming increasingly instrumental or performative, reducing the influence of humanist values and leading to a more quantitative approach that enables comparison against recent 'national averages' of performance based on standardised criteria. This has led to schools being described as risk averse, as they implement changes to their provision to ensure success in those areas that are critical measures of performance within the inspection framework, rather than their own interpretations of outstanding educational practices. The inspector is described as the conduit, gathering evidence of performance against standard measures, and using their professional judgement – a form of expertise in improving performance in core areas – to give advice on how to improve.

This literature review shows that the changes in inspection regimes across Europe and particularly in England, from more independent supportive mechanisms to a process aligned with government policy, and which attempts to influence school's 'output', follows a general standardisation or 'commodification' of education. This procedural approach, uses evidence bases of quantitative and qualitative data, gathered in objective groups under headings linked to presumed influencing factors for success (leadership, teaching, behaviour, etc.). The process of labelling and grading variables has led to a narrowing of focus, and a standardisation of the language of reporting, which has had a subsequent influence on the language and actions within schools. As reports are increasingly standardised, and the language condensed, this raises the question of whether the influence in schools will become equally restricted, or whether reports sustain their levels of influence despite the homogenisation of the process and terminology. This research attempts to identify language variables, and to follow presumed influence of terminology onto associated impact on children via the standardised mechanisms for measuring success built into the English school system.

2.4 Research Questions

The study has been designed to investigate the following questions, which began as general areas for exploration and became more refined during my discussions with school leaders and colleagues following an initial pilot project focused on the question 'How unique are Ofsted reports?'. On the basis of the quantitative findings in the pilot study, derived from the use of anti-plagiarism software, Turnitin, further hypotheses and qualitative and impact-related questions began to emerge.

Research question 1 (RQ1): How unique are Ofsted reports?

Many schools I have worked with describe reading very similar phrases and terms in Ofsted reports. Indeed, there is a widely held opinion among them that reports in their sector are overly similar. As many school leaders only have the time to read a limited number of reports, it is possible this is a skewed perception. Leaders tend to read reports from their own school and their main local competitors, who often have very similar contexts and cohorts. It is possible, then, that some of the same assessments or judgements, and therefore phrases, may be applicable locally, leading to a perception that reports are more similar than they truly are across the country as a whole. There is also a perception among school leaders that Section 8 inspections that convert into full Section 5 inspections have even more generic wording, as the time on site is switched at short notice part way through the inspection, which means the time to prepare activities and investigate is reduced. To investigate this on a large scale, software is used in this study that can, in concrete terms, measure the similarity of reports to give a definitive answer to the question of how unique the reports within this corpus are. Although the corpus is limited to a single calendar year, this is sufficiently large (1391 reports) to generalise to reports in at least the primary sector and to either confirm or challenge school leaders' perceptions of the similarity of reports.

Research question 2 (RQ2): Does the language used in Ofsted reports lead to subsequent improvements in outcomes?

There are some aspects of Ofsted reports that school leaders recognise as being more urgent than others to address. For example, if safeguarding is mentioned as a weakness in the report, action must be taken immediately and progress will be tracked vigorously not only by school leaders but also parents, governors, the media, and the Local Authority (LA). Many schools rely on the language used within the report to guide them in their identification of improvement actions, and specifically those that are stated as required actions within the areas for improvement section. The language of the reports is to some extent mandated by the requirements of the reporting schedule, (Ofsted Handbook, 2018) as inspectors must report on certain elements of school practice. There are opportunities for free text, however, and inspectors' choices of words to describe what to improve (for example, whether they say 'maths' or 'numeracy') will affect the subsequent actions leaders take in schools. There are no current mechanisms to link the impact of using different terms to specific improvements, and therefore it is not possible to determine which of those terms is most likely to lead to improved outcomes for children. Whether there are a range of terms that have similarly strong levels of power or influence on school leaders' actions and subsequent outcomes for pupils, and whether there is a scale of impact of terms that lead to improvements can, however, be identified by using software that can detect word frequency and usage and linking this to outcomes data for these same schools in those subjects for which there are published data. This research seeks to identify whether there is any correlation between the use of terms and outcomes for children. The multiple factors that influence outcomes will limit the findings here to general correlation rather than direct causality.

Research question 3 (RQ3): Are there trends within the language used in Ofsted reports?

Anecdotally, school leaders say that there are ‘fashions’ of language that seem to appear in reports in response to, e.g., new framework updates, launches of funding streams or research projects, which then seem to become a focus for school performance. Using its capacity to track words that emerge in or disappear from reports over time, the NVivo software can show whether there are any fashions or trends in terms that enter or exit common parlance within the corpus. This investigation will try to pinpoint trends within terminology used by the inspectorate and identify whether there are any terms that are particularly impactful, or whether linguistic trends have no correlation to other variables. Leaders often adopt the language of reporting within schools, from ‘triangulation’ to ‘deep dives’ (Ofsted, 2020), and the language trends of inspection have, in my experience, permeated school documentation and activity. Therefore, the identification of trends within reports could indicate likely trends within schools in the future.

Research question 4 (RQ4): Are there influencing factors that affect language or impact?

The study aims to identify trends in the language used across authors, settings, or other subsets of variables. If there is terminology that is used, e.g., by female inspectors, in large schools, or in outstanding settings, in particular, then this can be combined with outcomes data to see whether the formulation of advice, guidance or descriptions of some settings have a different impact on performance than others. This would also show any bias, conscious or unconscious, and could indicate why leaders tend to think that reports are so similar. If there are internal or external factors that influence language choices, and these choices affect the impact of reports on school improvement, this is critical for school leaders, school improvement officers, and inspectors to know.

Research question 5 (RQ5): Within the area for improvement, are there any individual terms that are particularly effective or impactful?

If a group of similar report contents, individual terms or groups of terms can be identified, and these can be linked to outcomes over time, concrete data on performance can be linked to these variables, to see if there are any variances in efficacy or components that are more likely to lead to improvement. Patterns could indicate that some terms are more effective than others. This kind of large-scale data pattern correspondence is one of the main uses of ‘Big Data’ within education (Sin and Muthu, 2015) and the large scale can mitigate to some degree the number of additional variables within the system. If there are terms that could potentially link to better or worse outcomes, then it would be useful for leaders, school improvement officers and inspectors to know the potential impact of their language choices on outcomes for pupils.

Chapter Three

Methodology

The design of this research takes into consideration the evidenced-based approach to inspection across Europe, and the existing research base that shows some tentative links between the process of inspection and subsequent improvement in schools. It assumes that there is an association between indirect intervention (reporting) and action taken within schools, and that there is a link between the choice of terminology used by inspectorates and national education policy. It accepts that the report is a 'tool for improvement', as stated in the 2010 Education Act and treats it as a critical component of the inspection and improvement cycle. As the inspector has some degree of autonomy when writing the report, and the report can include qualitative statements and language outside of the discourse of compliance and utilitarian performativity, the qualitative interpretation of data has equal weighting in the analysis undertaken.

Positionality

After many years of teaching and leadership in schools, I spent almost a decade as a school inspector for Ofsted, including time as an inspector for Estyn, the Welsh school inspectorate. Following this, I was appointed as a Senior School Improvement Officer', overseeing a large group of primary schools. I regularly used the skills and training I had developed as an inspector to help school leaders to interpret their Ofsted report and shape action plans to improve provision, with a view to improved gradings on subsequent inspections. In this position, I was privileged to work with headteachers and senior leaders with varying levels of experience in schools in a wide range of contexts and with different levels of Ofsted grading. I noticed trends in the patterns of support provided following inspection and the conscious and unconscious influence of the inspectorate on leaders' actions and vocabulary. This research project emerged out of an interest to explore this further, and my position as someone working with schools, who has experienced the inspection and reporting process as both an inspector and school leader, does inform the design of the research and the subsequent qualitative interpretation of the emerging quantitative data.

Given this background, I consider myself an 'immersed researcher' (Fraenkel, Wallen and Hyun, 2015) as it would be impossible to remove my own experience and contextual factors from my interpretation of the data. The combination of both a qualitative and quantitative reading of the data, as set out in this chapter, is maintained throughout this project, and is informed by an awareness that there are multiple ways of 'knowing' and interpreting the evidence, which here is shaped and enhanced by participation within the context and practices of the education, specifically school, sector. The mixed methods approach adopted not only follows established techniques within the field of discourse analysis but also mirrors the methods of evidence collation, interpretation, analysis, and reporting used by the Ofsted inspectorate itself, meaning that the research mimics accepted practice within the sector in its assumptions.

Framework

The study adopts Hyatt's (2014) approach to critical discourse analysis to investigate "education policy texts, and the processes and motivations behind their articulations, grounded in considerations of relationships and flows between language, power and discourse" (p. 41). This approach acknowledges that analysis of this nature includes not only the systematic investigation of the quantitative summaries of language trends and patterns but also the interpretive elements relating to the "agents and actors in the realisation, construction and perception and relations of power" (p. 42). This includes the assumption that across multiple documents, language trends may emerge that can be identified, classified, and which, by association, can have a correlation with or be a contributory factor affecting associated variables. It follows Wimsatt and Beardley's (1946) notion of interpretive fallacy, i.e., when reading we form an interpretation that we assume to be common. In the process of making meaning from Ofsted reports, my being situated within schools gives a specific interpretive perspective that is contextual and situated rather than based purely on the text's linguistic content alone, which can bring out a richer meaning and significance from the reports. Hence, the quantitative elements of the research (counting frequency of words, etc.) were balanced by a qualitative approach, using a more humanist (Pilkington, 1991) or hermeneutic (Hirsch, 1967) set of assumptions, whereby the art of understanding and the art of deriving meanings from the texts are enriched by participation within the culture in which they were generated and their intended audience.

The methodological framework adopted here allows for both the quantitative measurement of similarity and frequency and a qualitative interpretation of data at a macro level. This mixed-method approach is common in similar studies (e.g., Petty et al, 2012), although most of these are within health research and focus on the links between written reports from medical staff, the impact on actions taken by patients, and the associated clinical outcomes (Iacobucci, 2018).

This follows Cresswell's (2009) principle of isolating and analysing identifiable attributes that may impact upon an outcome, found within documents that represent discourses of power and influence (those of the inspectorate, government policy) and those who are to receive and act on those documents (school leaders and school improvement officers). The analysis here draws on Fairclough's use of critical discourse analysis (1989) rather than discourse analysis in order to develop generalised commentary on the power and impact of the discourses and subsequent actions in socio-political context from across the large sample of documents. The mixed methods approach is specifically designed to enable a more thorough understanding than each method could achieve in isolation, and the ability to triangulate (Fraenkel, Wallen and Hyun, 2015), leading to a more operationally useful set of findings. This research draws upon CDA methods in that it utilises the core principles of power and discourse strategies but departs from traditional CDA in that it was more iterative, and highly dependent on professional contextual knowledge such as the importance of the word 'good' and is more informed by statistics and Big Data than historical models of discourse analysis. The limitations of this include being not accepted by traditional CDA proponents or being so disparate as to be seen as a Big Data approach, devoid of the search for power and influence. Here, the 'power' that is sought is not sociologically but educationally situated, and assumptions around the English education system, such as the imperative to action, based on Ofsted's guidance, and public pressure for performance are embedded within the reading of each set of data and terminologies. The critical component emerged through several iterations of the analysis using traditional discourse techniques, and the lines of power and control emerged through reflection on the impact of the terms within the reports, rather than the research starting out to be a critical discourse analysis at the start.

Underpinned by the assumptions of critical discourse analysis, the design attempts to unveil the ideologies and power relations in the production (report creation) and manifestation (impact on leaders' actions and thereby outcomes) of Ofsted reports. The Critical Higher Education Policy Discourse Analysis Framework is particularly suited to this kind of analysis as it adopts transdisciplinary orientation (Hyatt and Meraud, 2015, p. 5) and enables the viewing of the document from multiple angles (as an inspector, a school leader, an improvement officer) and looks at the longer-term impact of the texts as part of a wider discourse over time. This was used to reflect on terms and their disciplinary boundaries as well as interpretive representations (e.g., what 'numeracy' means to different audiences within education and over time).

As the project looks at large numbers of words and a significant body of literature (1391 reports) the method needed to include approaches that were known to work consistently over large data sets. 'Big Data' approaches, such as those developed by Dastjerdi (2016) and Hesse et al (2015), show how the multitude of information generated by educational institutions such as schools or inspectorates are rarely observed and investigated at scale. Researchers are adapting tools initially designed for learning analytics to support the transition of 'data into knowledge' (Sin and Muthu, 2015) within education. This involves interpreting trends from within large data sets, using user experience to set 'frames' through which data can be viewed as contextually dependent, yet individually relevant. In this research, the 'frame' is shaped by the research question 'how unique are Ofsted reports?' and the software tools are then deployed then enable the identification of common concrete variables and comparators. At each point of the iterative process, we remain epistemologically talking about a 'group' of data representing a variable, rather than a single data point within this research for qualitative evaluation purposes.

3.1 Introduction to the design

Critical discourse analysis (CDA) (Van Dijk, 2009; Fairclough, 1989) was selected, as this was the most consistently cited method used in related fields to find similar linguistic trends, and had clear links to technological programs, as it quickly became clear that due to the size of the corpus (over a thousand documents) a systematic and specific set of technological tools would be needed to manage the volume of data.

The design was heavily influenced by discourse analysis techniques from similar projects outside of education (e.g., Morris, 1971; Breeze, 2011), wherein the strengths of the quantitative statistical analysis tools, which provide consistency and clarity (see e.g., Fairclough, 1989) and generate quantitative output, are weighed against the benefits of qualitative analysis that can offer a richer interpretation. Hence, in this study, quantitative functions, such as the ability to count word frequency, rank ordering, and T-Tests, are used alongside qualitative elements, such as the interpretation of the relevance of inspectors' vocabulary choices, or the collation of similar terms into a single code (e.g., pupils/children/students), and are explained as critical steps within the evaluation process. These steps can be seen in isolation, so that they can be replicated if required, or adapted for alternative use.

Generating quantitative data from a large corpus of documents

In the design of the research project, a wide range of approaches to analysing a corpus of documents were surveyed (e.g., Billig, 2002; Fowler, 1985; Wang, 2015, Van Dijk, 2001), to ascertain how researchers in related fields had managed similarly large data sets, and the comparative strengths and weaknesses of different approaches.

In the first stage of this research, Turnitin, a plagiarism detection software ordinarily used to check student assignment submissions, was used. The software is able to check similarity levels of a document/submission against others uploaded to the system and content available online. For this research, however, I created a ‘closed’ group, which meant that only the corpus of Ofsted reports was compared and other Turnitin and external internet content was excluded. The data obtained would therefore reflect only those items within the closed set. Initial checks of the set using Turnitin were run to identify how similar the reports were to one another, to test whether head teachers' perceptions of report similarity were supported by the data and thereby a potential area for further exploration.

A similar smaller-scale comparison has been undertaken using Turnitin as part of a pilot study that preceded this research (Lawton, 2019). In the pilot, a closed group of primary school reports published in January 2017 were checked for similarity, and a high degree of similarity was found (an average of 68%). Turnitin did not have the functionality to compare according to different variables, however, so in order to compare the entire 2017 corpus, and to compare the similarity of reports according to particular variables, e.g., schools graded ‘Good’ reports, reports written by a single inspector, inspections led by women, percentage similarities were logged onto an external spreadsheet. These variables could then be logged, and the data filtered by these contextual factors. The pilot showed that the most unique section of the reports was the areas for improvement. A second full corpus was created containing only this section of each report. The percentage similarities of these were then logged as a second data strand in the spreadsheet. This way, percentage similarity trends of single variables could be easily and quickly filtered, and averages calculated against full reports and for the areas for improvement section only (see Appendix 1).

A spreadsheet was used to record contextual factors/variables of each report, e.g., the gender of the lead inspector, the size of the team, the size of the school, date of inspection, or the gradings for individual sections. These influencing factors were then logged against the school alongside the similarity score of a report, word frequency, outcomes, and judgements to identify language trends by groups, or the efficacy or language similarity of subsets of the corpus.

In order to undertake more advanced analysis, such as word frequency searches and sentiment coding, as is common within discourse analysis, the use of more advanced software was explored. NVivo was selected, based on its effective use in similar projects. NVivo is a recognised tool in discourse analysis (Trena, 2017), which has undergone many refinements since its launch. It can not only filter, but also its in-built algorithms enable researchers to interrogate linguistic data through multiple lenses. Using this, codes could be set up for specific contextual factors, gender of lead inspector, size of team, etc., so that data could be explored in context or to confirm or further investigate emerging trends from the initial spreadsheet. The software also had the ability to ‘auto-code’ by linguistic terminology similarity, saving thousands of work hours, and using an algorithm that is more consistent than human coding of English language texts, and which will automate the ‘stop word’ functionality so as not to introduce bias. Classifications were set up, such as month of publication, and used as filters, so that analysis of ‘word frequency’ by month could be undertaken. The information produced by this filtering was then used to investigate trends over time, across inspector classifications and for criteria such as ‘maintained’ compared to ‘academy’ status. These results were then analysed to identify trends and any available answer to the research questions.

During the year under investigation, 21% of all short inspections ‘converted’ from Section 8 to Section 5, i.e., began as a short inspection, but due to a potential change in grade became a full Section 5 inspection during the on-site visit, generating a full report. Half of these full inspections resulted in a drop in overall grade. Following a short Section 8 inspection, 90% of primary schools had stayed the same or had improved to Outstanding.

Another factor contributing to the choice of sample in both the pilot and full study was the grading profiles. According to the Ofsted statistical release for the corresponding period (Outcomes report, 2017), the proportion of primary schools in each grade boundary is relatively stable (Figure 4), which reflects a reasonable representation of the proportions over time. By contrast, there is a notable change in secondary schools during this period, which Ofsted attribute as largely due to ‘closures’ (106 schools closed in 2017, but this often refers to a change of status from maintained to academy school and therefore a change of Unique Reference Number (URN). At this time, many schools were converting to become academies, either by choice or due to being seen as ‘eligible for intervention’ due to a number of non-compliance or non-improvement reasons (otherwise referred to as ‘forced academisation’). In the process of academisation, schools leave the oversight of the Local Authority and become the responsibility of a chain or Multi-Academy Trust that is privately sponsored. This had a statistically significant impact on the number of schools and proportions of ‘Good’ schools overall. There was a decline in the number of secondary schools graded ‘good’ during this period, which is hidden by this method of reporting the data, as many ‘closed’ and reopened with a new name as an academy. The academisation of schools has an impact on data trends, due to the government requirement for underperforming schools to academize, which had affected secondary settings more than primary by this point in the first instance, with the impact at primary found in later data sets.

This could be due to higher prior gradings, or the influence of the church, who were able to reject or postpone proposed academisation of faith schools, of which there are significantly more primaries than secondaries.

Primary school reports had the most ‘similar’ patterns of content, as their reports followed a more consistent structure, whilst maintaining the scale of the sector. As there were only 133 secondary reports compared to 1678 primary reports in 2017 (Ofsted, 2018) these would have generated much smaller groups, but with the necessary filtering out of those with sixth forms (who have additional written sections), those too small to generate groups, and those with no data (academy conversions, closures, etc.), this would have reduced the corpus, skewed any word frequency trends, and resulted in groups of schools small enough to be identified, potentially creating ethical issues. Primary school reports are more numerous and conform to the report template more closely in terms of structure and required content. Even if those too small for comparison are removed, the corpus remains large enough for large scale data mining for word frequencies to provide a measure of linguistic trends.

A limitation of this approach is that small schools or those with non-standard contexts, such as Pupil Referral Units, specialist provision, and infant and junior settings cannot be easily included, and so some groups are not represented within the sample, and these include a skewed proportion of schools in coastal and rural settings.

3.2 NVivo Tool and power relations

The automated ‘stop words’ list was used to discount terms that were irrelevant or would skew the data. Hence, phone numbers, email and web addresses, the names of months, days, and years were added to the stop list so that these did not bias frequency counts, as each report states not only the month of publication on every page, but also references the dates of the last inspection.

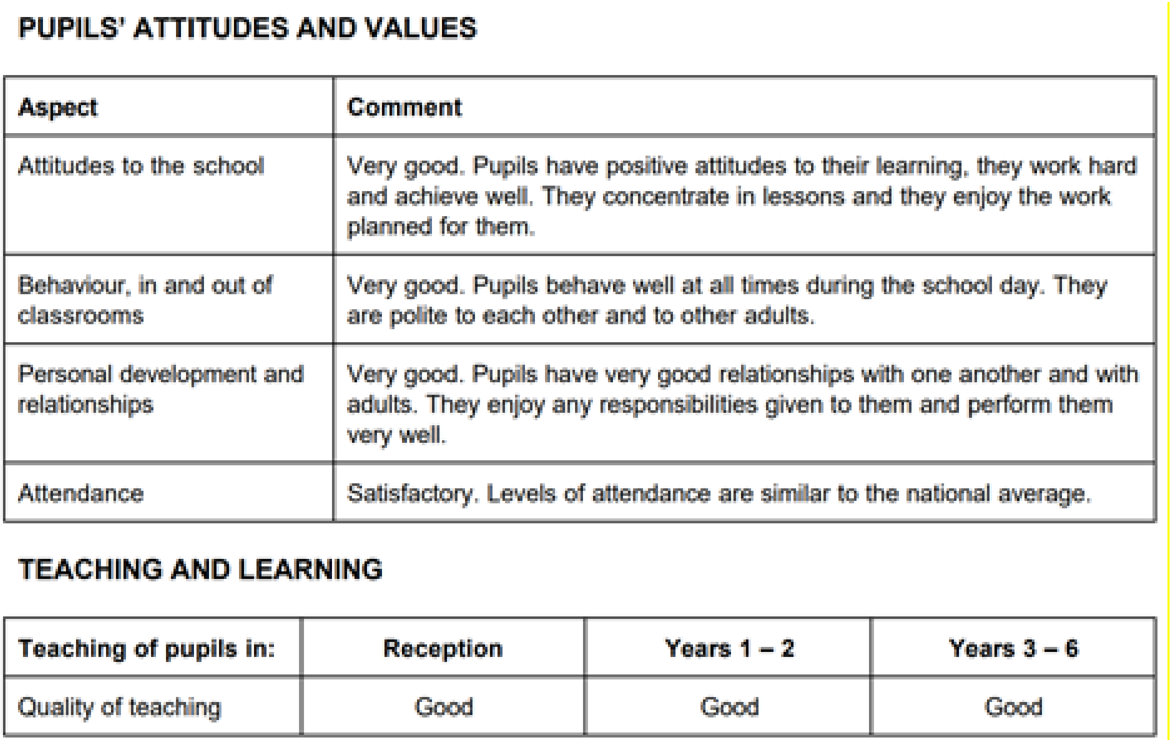
Critical discourse analysis (CDA) looks for power relationships within texts: who is speaking, who is reading, and what transfers of power or authority lie within the use of language and the text itself. As the reports are a legal document issued by a government agency to a school, as a tool of school improvement, and are also written to enable parents to make informed decisions about school choice, they have multiple uses and purposes, and layers of power within the language. One of the criteria for the quality assurance process undertaken by Ofsted before reports are published, is that the reports must have a low reading age. This can be checked using Microsoft Word using the ‘readability statistics’ function. For example, the ‘Leadership and Management’ section of inspection report

#1003678 scores a reading ‘ease’ of 42.1/100 (difficult to read). Reports normally aim for a minimum of 60/100 which has a reading age of roughly 15-16 years old. To reduce this reading age and increase ease, some key education terminology, such as ‘*differentiation*’, is removed from reports during the Quality Assurance (QA) process. This may lead to inspectors to use a limited vocabulary for reporting purposes from the outset. In this research, however, it is not the ease or reading age of the reports that are being investigated but whether this kind of limitation on language due to purpose and audience, alongside aspects such as word count limits or requirements to address specific core areas (e.g., progress of disadvantaged pupils, comparison with national data, etc.) limits the range of content of reports, thereby producing high levels of similarity, so much as to render them not useful or fit for purpose.

The specific details of the Ofsted quality assurance process is not publicly declared. In the 2015 Ofsted handbook, which would have been in use for the 2017 reports, the phrases “inspection reports will be quality assured before Ofsted sends a draft copy to the school” (#121) and “If Ofsted decides that a report should be subject to further quality assurance, the school will usually receive an electronic version of the final report within 23 working days” (#125) are the only details relating to the quality assurance process. From my former roles, as both a lead inspector and senior managing inspector, I have knowledge of the internal quality assurance processes in place during this time, but confidentiality agreements preclude my detailing their specific content. As such, the focus of the research is on the documents in their final form, which are unrestricted and in the public domain, and not on the details of the quality assurance processes that may have influenced vocabulary choices.

Software has been chosen that can identify what words are being used, and how frequently. NVivo is often used to show patterns of language, ranking words by frequency of use, identifying strings of terminology and can identify synonyms to group terms in multiple extraction patterns so that trends can be identified. These patterns can also be investigated in terms of any single variable or combination of variables, e.g., the author (gender, inspector designation), contextual factors (team size, school designation), inspection pressure (previously good, first inspection). These factors, by using the software tools of NVivo and Turnitin in combination, can also be investigated for similarity, so that dependent variables can be seen. This approach to finding correlations constitutes a ‘Big Data’ systematic interrogation of a large data set (Dastjerdi, 2016), used to find meta-patterns that are undetectable by an individual’s reading of a single or a small group of reports, as would normally be the case in school leaders’ professional practice.

It could be that the limitations of the word count, the structure, and the inspection criteria mean that it is impossible to capture the unique characteristics of a school in an Ofsted report. Given this, reports could, in the future, consist of coding according to a set of descriptor indicators. For example: "School X is Good. Positive for descriptors 2, 3, 4, and 10, needs to improve 1, 5 and 9". The inspection process would be more iterative and less bespoke, and much closer to earlier permutations of Ofsted reporting, where gradings for individual statements or lesson observation summaries were included (Figure 2).



PUPILS' ATTITUDES AND VALUES

Aspect	Comment
Attitudes to the school	Very good. Pupils have positive attitudes to their learning, they work hard and achieve well. They concentrate in lessons and they enjoy the work planned for them.
Behaviour, in and out of classrooms	Very good. Pupils behave well at all times during the school day. They are polite to each other and to other adults.
Personal development and relationships	Very good. Pupils have very good relationships with one another and with adults. They enjoy any responsibilities given to them and perform them very well.
Attendance	Satisfactory. Levels of attendance are similar to the national average.

TEACHING AND LEARNING

Teaching of pupils in:	Reception	Years 1 – 2	Years 3 – 6
Quality of teaching	Good	Good	Good

Figure 2 Extract from an older style inspection report #247845 (November 2002)

This kind of reporting still exists in the independent sector and Welsh and Scottish systems. This would imply that there are standard actions that can be undertaken in every setting, regardless of context, which would result in a grading of 'good'.

This would mean that anyone reading the report is not getting a narrative, but an audit, something the new framework was initially designed to move away from. That a longer, narrative report is produced in the current system of inspection implies a greater level of bespoke advice and tailoring to context than an audit design. In 2022, the DFE is moving towards a more audit-centric system, with learning analytics gathered through academy reporting structures, more centralised control of testing and performance against financial targets (schools benchmarking website) widely accepted mechanisms for measuring performance. The White Paper released in March 2022 (Opportunity for all: strong schools with great teachers for your child, DFE), the first paper released in six years, indicated more quantitative measures of success, including target percentage scores in both SATS and GCSE.

In similar projects, single and combined variables have been investigated to identify which variables have an influence on outcomes (Petty et al, 2012) In this project, report similarity or inspector designation could be used as the dependent variable. This brings the project design in line with similar investigations. Corresponding approaches to 'discounting' were used, where outliers to the core corpus (i.e., those reports with unusually amended content, due to very small size or unusual context) were rejected as not pertinent to the trend analysis. Dastjerdi (2016) describes the filtering of outliers as a necessary part of data capture when working at this scale, as single data points have less significance within the larger data set, and the strength of the data lies in the correlation and patterns at the macro level. This pushes for a more quantitative approach, and the associated suggestion is that direct causality is not as important a finding as trends at this level, as the influence or response is described at the group rather than individual level, and generalisations refer to the group rather than any individual member.

This enables a particular separation between the quantitative, replicable 'data' and the interpretation of those findings. The concern with the operation of power lies outside of the numerical cataloguing and emerges within inference of what or who has led to those patterns, and what those patterns mean for those receiving the reports. The patterns themselves remain solely mathematical representations of correlations and trends, or not. Labelling which trends (or lack thereof) are important or meaningful to the primary sector, or education sector as a whole, are shaped by my own professional experience, knowledge, and interpretation of the data.

3.3 Software assumptions

'Computer assisted qualitative data analysis software' (CAQDAS) such as NVivo, does not automate analysis, but produces summaries, undertakes searches and follows algorithms that have been requested, designed, and run by the researcher (Woolf and Silver, 2017). Just as Microsoft Word cannot write an essay independently, NVivo cannot undertake searches and find meaning in their outcome without researcher involvement. Assumptions built into the software linked to the English language - plurals, sentence structure, punctuation, and so on - are common across linguistic research (Trena et al, 2017). The automated counting of words, or logging phrases as 'positive' or 'negative' based on vocabulary choice, is now established practice in managing large corpus, that historically would have been manually completed. These coded themes can be identified, recognised, and are included within the method in Chapter 3.2 for clarity.

The Section 5 reports on primary school inspections published from 01/01/2017 to 31/12/2017 were identified using the Ofsted report online search tool. An initial sift of the reports was then run using Turnitin to identify any reports relating to middle schools, alternative provision, or any setting other than a primary school. These were not included in the sample as the reporting template at the time differed for different settings, so these are not comparable with the majority. For example, reports on those settings without Early Year Foundation Stage (EYFS) provision had notably less required content. Those who received a Section 8 letter, or report outside of these parameters, were excluded as their content was also not comparable.

Once the corpus of reports on primary schools that had undergone a full Section 5 inspection were identified, infant and junior schools were discounted, as their reports also had differing content: junior schools had no Early Years grading or commentary; infant schools had no Key Stage 2 outcomes. This approach follows standard discourse analysis techniques (Van Dijk, 1993) used to reduce the language variables to ensure that any variance found later was meaningful within the parameters of the same specific report format. This enables the study to focus on the core vocabulary of one standard format of report rather than incorporating differences in style or content requirements, as the software used would have highlighted these reports as outliers, and their inclusion would have skewed the word frequencies. Settings where there were fewer than 30 pupils were also discounted, as these follow a different reporting template and have differing terminology restrictions. For example, if there are fewer than five children in any core group (pupil premium, boys, etc.) then the report avoids evaluating their performance against these, in case individual children become identifiable.

Following these exclusions, the final corpus consists of 1391 primary school full Section 5 reports published in 2017. These were uploaded to Turnitin and NVivo, and the spreadsheet of key variables was populated. All reports included from within that year had had the full Ofsted reporting framework applied, and all were in the public domain. Lead inspector status, gender, date, and team size were labelled and analysed for trends. Through this labelling, a small number of inspections were identified that had been interrupted and restarted. Some had multiple dates of inspection periods, others had two lead inspectors named. These reports were included, as a single report was generated, despite multiple inspection activity points. In practice, this can be due to schools closing unexpectedly, e.g.

due to snow, or because an inspection has been deemed ‘incomplete’ during the quality assurance process. The QA team will deploy additional inspectors to collect any missing evidence (Ofsted Handbook, 2017) to consolidate judgements (rare), or even a full second team to check the accuracy of a contested judgement (very rare). It would be impractical to ascertain the circumstances of these cases without further investigation, and this additional detail is unnecessary for the purposes of this study. If any of these reports had flagged as unusual, then this could have been further investigated, but this was not the case. This is an important finding, as despite having two different leads/creators, reports conformed to the same patterns of replication and similarity as a single author.

The corpus was analysed using Turnitin to identify similar or duplicated phrases across this closed set. Similarity reports were generated that identified the location of almost identical phrases, mirroring the data mining principles used within NVivo. Phrases are identified, labelled as a code and a list generated of all uses of this term, showing how many times it is used, and specifically where within the document. The similarity report shows what percentage of an Ofsted report is identical or almost identical to another report, which terminology are the same, and which report it duplicates. It repeats this for any duplications, for as many sources as are found. For example, one report was 54% similar to another within the corpus. The extract below shows how the Turnitin report identifies similarity – highlighting the words that are the same and leaving in another font or colour those that are different. In this case, only the words ‘…and writing’, ‘English’, and ‘Consequently’ were different, with the rest of the paragraph identical in both reports.

> ***“Teachers do not consistently challenge pupils in their learning, particularly in mathematics*** *and writing*. ***As a result, pupils, including the most able, do not make as much progress as they should. Teachers do not consistently reinforce high expectations for pupils’*** *English* ***spelling, grammar, and punctuation across the curriculum. There are too many inconsistencies in the way mathematics is taught across the school.*** *Consequently,* ***pupils' progress in mathematics is not improving rapidly.”***

Extract from a Turnitin similarity report within the corpus.

Similarity percentage reports were generated for each report in the corpus. On analysis, it became clear that the main body of the reports followed consistent patterns, and the most unique part of the reports was the areas for improvement section. As these could not be easily extracted for separate analysis within the Turnitin software, a second corpus was generated, using only the areas for improvement section of each report, so that the similarity of only this key section could be compared as a closed corpus. These similarity percentages, and the highest single source for each report were logged in the spreadsheet. Using the data in this way, the similarity profiles of different sets could be identified by filtering the spreadsheet.

These initial trend patterns were used to narrow the variables used for further scrutiny to identify whether any single or groups of variables had distinct trends and differences. This followed established approaches to interrogating linguistic data at this scale (Van Dijk, 2001), isolating and investigating variables as single elements of a complex context. The aim is not to identify a single causative or correlating factor, but to see a trend that might be otherwise unseen within a large body of texts written by multiple authors. During this first stage the focus was on the 'input' element (e.g., authorship, setting), which are concrete, quantifiable variables, in order to identify trends independently. Focus on the 'output' requires qualitative interpretation and forms a second stage.

Once the reports were sorted and coded on NVivo according to input categories, word frequency analysis was run to find the most frequent 200 words, which were then logged into a spreadsheet. This showed that variation in frequency of single words was negligible across full reports under each of the variables (lead inspector status, gender, etc). The most frequent 200 words were used as this included the right-hand side of the bell-curve of frequency distribution of words, from the most frequent down to roughly the mean for each list. Below this point words were so infrequent as to not generate comparable trends.

The NVivo 'auto-code' was then run on the full reports data set, to see if the system's algorithms identified any trends, and sentiment analysis ran on the same data to see if any trends in positive or negative language could be ascertained. The two processes use similar algorithms to code the reports into groups. The internal 'auto code' system identifies similar phrases and sentiments by grouping nouns and synonyms and identifies whether the phrase as a whole is positive or negative. For example:

The outcomes for… [pupils/students/children] are… [good/strong/excellent or weak/poor/low]

This creates overarching 'code themes' e.g., all text relating to 'pupils' or all text referring to 'English'*.* These internal algorithms were used to run analysis at 'paragraphs' (bullet points of text) rather than 'sentence level' in order to mimic the writing style of the report. Sentiment analysis labels phrases as 'very positive', 'moderately positive', 'moderately negative', or 'very negative'. This then provides a further possible criterion by which to filter the reports, to see, for example, how many of the phrases within a report are ‘very negative’ and include ‘pupil’. The sentiment analysis is based on English language vocabulary choice for adjectives, etc. and is not influenced by the individual researcher, so although a level of interpretation is involved, it is syntactic rather than contextual. Each sentence is coded in isolation, and the level of sentiment is graded solely based on the terminology used.

Use of NVivo’s auto-code function created a list of themes, which could be logged as individual codes and interrogated as groups, or by code to determine frequency. For example, two reports #117342 and #101328 both had the fewest codes (52), most had between 61 and 100, the three with the most codes were #111531 (173 codes), #108699 (184 codes) and #140873 (186 codes).

Initial variables investigated included word frequencies and sentiment analysis:
-by month
- by gender of lead inspector
- by whether HMI/OI led the inspection
- by school faith designation
- by team size
- by overall grade category

Through this process, although few themes related to difference by labelled variable emerged, the process did identify the most commonly referred to curriculum subject areas (Mathematics; see figure 6), and the most unique area of the report. The process used established linguistic principles for frequency, such as the inclusion of ‘stem’ and plural forms, tenses as well as apostrophe permutations to identify the most frequent subject area. In order to further investigate the impact of the report on schools, I linked the two findings, most frequent subject area and most unique comments, which enabled me to follow a trail from research question through to quantitative outcome. Given that mathematics was repeatedly given as a required area to improve, with varying degrees of similar and bespoke language, and the published mathematics data from those schools who were given mathematics as an area to improve were able to be tracked over time for these schools, I was able to see if the result of schools being required to improve mathematics using different descriptive terms did result in improve mathematics outcomes for children, as shown in externally verified data.

This stage required more complex management of data, and a second spreadsheet was created, including the URN, published SAT mathematics scores for 2016, 2017, 2018 and

2019 by percentage of pupils reaching age related expectations and average point scores. These years were chosen as they would represent the school's performance in mathematics prior to the inspection, at the point of inspection, after 12 months of improvement, and roughly the point of re-inspection for most settings, at which point progress against the Areas for Improvement area would be checked. These specific mathematics data points were used as they were a consistent part of mandatory publication requirements for schools during the same period.

Mathematics is a concrete data set, with a comparable national performance indicator (Ofsted, Statistical Data Release, 2018) and is a required reporting feature of all schools, including those the initial sample of reports represent. During this process of collating mathematics data, some schools had to be discounted. Following inspection, some had closed, or became academies, and so did not publish data for some or all of the time period. Some were very small schools and had less than 5 children taking the SATs (Standard Assessment Tests), which requires that no data for those schools are published in case individual children are identifiable. Another group of schools had published data for some but not all years during the period, and were discounted as unreliable for trend analysis, as numbers of KS2 children, and thereby data, had varied significantly from normal distribution.

This meant that 54 schools out of the 634 in the corpus for whom mathematics was identified as an area for improvement were removed (see Appendix 1). This affected the represented proportionality of small schools, who are generally inspected by a single inspector, and more usually by an HMI (which is borne out by the full data set). The variance is a maximum of 4% on single inspector small school inspections, which is included in any analysis of the data. All other areas are within 2%, which is reasonable tolerance for a data set of this size (Dastjerdi, 2016).

3.4 Discounted lines of enquiry

From the initial spreadsheet, categories from the core variables identified from the reports (lead inspector status, team size, etc.) were logged alongside the Turnitin statistical similarity figures, for both the full reports and area for improvement sections alone. Although the type of setting was logged (e.g., maintained, academy, free school, community school and so on), this proved to be a redundant area for exploration, as the grouping proportions in all variables roughly reflected school types nationally (Ofsted, Statistical Release, 2018) and no group showed any spike in grading, gender of lead, team size or other variable. These categorisations were saved to enable further exploration of linguistic trends emerging from the initial auto-coding and data sorts relating to 'author'. Faith designations of schools were also tracked, but no trends were found. Searches of specific early years terminology using NVIVO word frequency showed no obvious patterns, and this line of enquiry was abandoned.

The reports' geographical areas were also labelled in the first stage, so that any possible north/south trends or regional variations in language could be ascertained. Reports include the postcode of the school, so this data was added to the spreadsheet. This addition rapidly became too big a workload compared to the other variables and was abandoned as a line of enquiry. Beyond postcodes, the range of ways of dividing the country and comparing results would have required further investigation into how inspectors are allocated, regional dialect differences, and so on. It remains an interesting potential line of enquiry for future research, especially since 2019, when inspectors now work in defined regional areas, which would give reasonable inspection-related reasons for specific geographical groupings.

3.5 Analysis of the research model used

The quantitative analysis of the data was carried out using statistical approaches, including T-Tests for comparability of means, Dunnett's tests as post-hoc, and Anova where there are comparable groupings. All T-tests used are 2 tailed, 2 sample equal variance (homoscedastic) tests, using the common assumption of P< as 0.05 as significant. This is because means were often groups of different sizes, and comparisons needed to be made between 'with' or 'without' a certain word (for example, those reports that did use the term 'mathematics' and those that did not), and between the performance of one particular variable compared to 'all' (for example, comparing the performance in schools where the report used the term 'problem solving' to the national average, or to the performance of those schools that had a mathematical area for improvement).

The identification of variables, translation of data into comparable lists, and use of the built-in functions for T-tests and Anova within Excel spreadsheets improved the speed and reliability of results. The statistical significance was taken as evidence of the relevance of the line of enquiry, but judgements were also made by the researcher as to their relevance to practice, based on professional knowledge and experience, for example, which words were newly added to the SATS test in recent test cycles. There is a risk that experience could also skew the choices made at this stage, and the parameters of each decision made, its basis in professional experience and how it relates to the original research questions, is explained in what follows. During the analysis of the data, several lines of enquiry were begun and abandoned due to the lack of clear results, or no correlations being found. These are included to aid transparency. The iterative process revealed further questions, and these are included as possible further lines of enquiry throughout the discussion.

3.6 Ethics

Consideration was given to the full range of BERA ethical principles, and the identification of individuals affected by the research. Although some inspectors could potentially be identified, as their names are associated with individual reports, some of which are quoted here, the regulations for publication and ownership of the report by Ofsted as an organisation rather than by an individual is an important factor. No privacy, autonomy or individual values are investigated or evaluated, and all findings relate to the system rather than any individual.

The analytic tools used are common in the sector and similar sectors, and although their use and the specific combination of tools is unique, their use follows principles of data handling and are open to interpretation by multiple observers. The design was set up to maximise the benefit not only for the project, but also for future use of the data and interpretation of the findings by other researchers.

The documents included in this study were already in the public domain and had already been through a stringent quality assurance and checking system by the inspectorate's own internal processes prior to publication. Trends, themes, and patterns are only discernible at the macro stage, once individual reports have become part of a much larger corpus. Individual inspectors and school leaders are named on the reports, but these are not referenced by the data sorting process, and no groups are so small as to render individuals or schools identifiable. This was confirmed during the initial filtering processes. Filtering by name of inspector was rejected to limit the impact on individuals, but this is freely available public information, linked to the reports by Ofsted themselves, and could be tracked. Gender was used as a variable where the inspector could be identified by pronouns on the inspector database or within the report itself, which for this corpus happened to be the full data set.

Any conclusions drawn or recommendations made here relate to the system of reporting, the use of the data processing and analysis tools, and the impact of the 'areas for improvement' sections of the reports as instruments for school improvement, rather than to individual inspectors or schools. All data used in second wave processing, SATs (Standard Assessment Test) results, current inspection status, or performance data is also all in the public domain and, where possible, was averaged out as mean scores. Those data sets that are shown to relate to individual schools only show publicly available information. The corpus does include schools with which I have been professionally involved, as an inspector and/or School Improvement Officer, without fear or favour. Any trends identified or conclusions drawn bring no detriment to my own working life, relationships with the inspectorate, or with the schools. No politically contentious outcomes have been found that could be damaging to Ofsted, to myself, or to the schools within the sample.

As the corpus under investigation includes reports that I have written and published as lead inspector and includes reports of schools with which I worked as an improvement officer, consideration was given to the impact on my own practice and professional life. No findings from this study are detrimental to my work with any specific school or those schools I have worked with or inspected in the past, as all schools' reports are treated equitably and proportionally as part of the corpus. No reports or schools with which I was involved were outliers within any data set. During quantitative data gathering, individual schools and reports were anonymised, so that neither myself nor the software were able to identify those schools I had personally worked with during the analysis phases in order to reduce the possibility of bias.

Chapter Four

Empirical Findings

Within Chapter 4, how the data set was refined and analysed is laid out. Initial sorting and sifting are explained, and the first wave of NVivo statistics relating to word frequency and how this influenced further analysis is discussed. In 4.2, the similarity data is explored, and the three main patterns exemplified. 4.3 describes how the quantitative and qualitative elements were used to refine further investigation of the research questions, and 4.4, 4.5 and 4.6 are cases where a single variable was investigated using both methods to explore potential findings relating to the research questions. 4.6 and 4.7 are particular cases where the literature review suggested there may be findings that could be considered against similar research in the field. The core findings of this project, relating to the 'area for improvement' section of reports is explored in 4.8, and the link between reports and outcomes for pupils tracked in 4.9.

Initial Data Set – setting up the corpus

The overarching data on inspections during the period of study was readily available from the Ofsted inspection data website, within the document 'Maintained schools and academies inspections and outcomes as at 31 December 2017'. This showed the judgements for each sector and number of inspections that year against the total number inspected, and breakdowns of inspection type and geographical area. From this, I was able to evaluate the corpus as roughly representative of the inspections in that sector for that year proportionally compared to the sector, the grade profile, and geographical spread (Appendix 1).

Compared to 'All inspections', the corpus is skewed to more reports on schools awarded Grade 3 ('Requires Improvement'), with fewer Grade 2 ('Good') schools. This is because, in the period covered, all Grade 3 schools received a Section 5 inspection, which generates a full report, and which was therefore included as a comparable document as discussed in the previous chapter. During this period, a proportion of schools previously judged 'Good' or 'Outstanding' received a 'short' Section 8 inspection, which generates a letter rather than a full report. These letters are not included, as they do not follow the structural or linguistic conventions of a full report. Also, these letters have less impact on school improvement choices, and therefore outcomes, due to their 'interim' nature, and as such are less relevant to this project.

4.1 Initial NVivo statistics

Once the corpus had been established and analysed to confirm proportionate representation to answer RQ1) How unique are Ofsted reports? the NVivo tools were then used to run an initial analysis to investigate RQ3) are there any trends in the language used in inspection reports? As NVivo offered a range of approaches to investigate trends within the language, based on discourse analysis techniques from similar studies within the literature review (see chapter 2.2) the approach taken started with the larger, more generic investigations of patterns and used an iterative process, informed by professional practice to refine down to word-level explorations. This is common within large corpus, using coding and referential triangulation (Fairclough 1989) to compare similar, thematic, and oppositional phrases, trends, and patterns.

Sentiment analysis

The analysis started with a sentiment analysis (QSR definition, 2012) of the entire corpus, at 'paragraph' level using the software's internal algorithms to identify positive or negative sentiments (see Appendix 8). This choice was made because the report writing style is naturally broken down into short paragraphs. This analysis gave a result of mixed positive and negative sentiments across all reports within only one not gaining any 'very positive' and 45 schools awarded no 'very negative' logs. For example, terms such as 'exciting' would be graded positive, whereas 'struggling' would be classified as negative, and signifiers such as 'very new and exciting' or 'significantly struggling in many areas' would indicate potential for grading 'very positive' or 'very negative'. This was interesting, as although the proportion of 'inadequate' schools was small (15 schools), only one of these did not have any 'very positive' comments. Equally, there were 162 graded 'outstanding' schools, and all of these had at least one 'moderately negative' comment. 97% of the corpus had 'very negative' comments logged within their report. Each of the sentiments, very positive, moderately positive, moderately negative, and very negative were coded as sets and run through the word frequency tool to see which terms were the most common for each sentiment (Appendix 9).

This list of terms was then investigated for trends or patterns. Sentiment analysis had generated large numbers of coded sections (85,781) which meant that only sizable variances would show as statistically significant. Alternative routes into this data, for example, coding only the 'area for improvement' for sentiment, or coding only the front page may have led to more efficient routes to word level findings. However, the process remained iterative at this point, and instead, this data was used to shape word frequency queries, for example, as highlighted in the table in Appendix 10, '*teaching*' features highly in the most frequent comments for both positive and negative sentiments, whereas far fewer very positive statements about '*safeguarding*' were noted, compared to more very negative statements about safeguarding. Similarly, the term *'disadvantaged'* is linked to more negative statements than positive. Some terms, specifically related to the actions of leaders (*regulates, inspects, services, concerns, guidance*) occur only in the most negative terms and not the positive. The inference from this is that if the 'area for improvement' section of a report accurately reflects the report content, then more actions relating to leaders should be found as areas to improve than other aspects.

Word frequency

A word frequency search was then run on the full corpus, using the automatically designated stop words from NVIVO for English language, and adding to that stop list phone numbers, web addresses, months and page numbers added as these are common to every report. The search was limited to 'most frequent 200 words', as the sentiment analysis had shown that frequency drops substantially after this point, limiting significance and identifying individual schools (for example, the word 'farm' spikes for 5 schools who have farm in their name). Stemmed words (e.g.: teach, teacher, taught, teaching) were used rather than exact matches to limit the variables in the results.

This initial word frequency report (Appendix 10) was then used as the comparator for all further investigations, so that any spikes in different groups could be compared against the original full corpus frequency report. This gave a baseline against which other data mining could be compared, and the ability to group words into themes for possible trends and patterns, and mimics the Turnitin approach, giving consistency to the data analysis. These highlighted additional words to add to the 'Stop List' (e.g., the word 'inspector') and the grouped table (Appendix 11) was created to code the frequent terms, with groups informed by practitioner experience. From those that remained, some were coded 'other', and some were flagged as possible stop words, as they were non-content related terms (e.g., '*also*'). This table was saved so that if these terms spiked, they could be investigated, discounted, or added to the stop list.

Critique of the approach

As the frequency and sentiment results had proven to be a reasonable proxy for human coding using this format, the automatic 'auto code' algorithm was run to see if that could be used to easily access trends within the language. Similar discourse analysis techniques (Nunez-Perucha, 2014; Lazar, 2005) have shown some strengths and weaknesses of the system (Sakr, 2016; Martin 2004). These include criticism of the disconnected nature of word-level filtering from inference and how contextual language is easily mis-interpreted by data mining in this way. In response, the data was initially filtered using the automatic processes and then contextualised by qualitative analysis of the terms within bullet-level phrases, reflecting on purpose and meaning before evaluations were reached. In practice, in schools receiving an inspection report, it is a combination of the words themselves, which have a culturally specific meaning, for example the word '*safeguarding'* implies a culturally understood series of rules and regulations. There are also contextual and inspection-specific terminology, for example, '*embed…'* is often used to describe things that were seen during inspection, but which neither the inspector nor school leaders could be certain would continue long term.

The NVivo system auto coded into groups that correspond to the word groupings:

> Descriptors: *Development, Effective, Improvement, Needs, Funding*
> Education Terms: *Disadvantage, Premium, Progress, Information, Skills*
> Adults: *Governance, Leadership, Teaching, Support, Staff*
> Children: *Children, Pupils, Students*
> Settings: *School, Local Authority*

This enabled the word frequencies to be interrogated alongside those settings with most positive, most negative, or grouped by terms to see if there were any trends by spreadsheet variables (gender of lead, size of team, date of inspection, etc. See appendix 1). No frequency trends within sentiment were noted for any groups or variables.

This data set showed that the most common terms were prolific across all reports and to such an extent that the original complaints of head teachers that reports were 'overly generic and similar' is a reasonable interpretation. The most frequent 200 words occurred excessively in all reports to an extent that all were in both negative and positive sentiment equitably, showing the terminology used was often almost identical, with only positive or negative signifiers separating the reports:

> "The pupils/the staff [have / have not] … Leaders [always/often/sometimes/rarely]"

Therefore, it is critical in analysing these texts to interrogate the data at word level and context level at the same time so as to discern any trends or patterns which are beyond this structural difference, and to investigate why the terminology was so similar and whether that had any impact on the power or influence of the report through RQ3 whether there are trends within the language of the report and particularly when going further into RQ5 'within the area for improvement, are there any individual terms that are particularly effective or impactful?' as single words within the area for improvement would hold more power, as word count is limited, and signifiers therefore critical.

4.2 Turnitin similarity data

To further investigate the corpus, the similarity data from the Turnitin process was analysed for trends, patterns, and variables. This was to move from RQ1 how similar are Ofsted reports? To RQ3 are there any trends within the language of the reports?

Initial Turnitin data showed that the spread of similarity was a rough bell curve (Appendix 3), with some outliers of very similar (88%) and unique (31%) reports, with most reports being between 43% and 68% similar. The average was 55.6% with a mode of 50%, a median of 54%. In summary, for this corpus, each Ofsted report is likely to be comprised of roughly half content that is duplicated from other reports within the same year. As a point of comparison, any student essay submitted for examination that is more than 22% similar to another student essay would be automatically flagged by this software as potential plagiarism. From experience as both an inspector and a school leader, some duplication is anticipated as reports condense broad themes and many schools face similar difficulties and are celebrated for similar strengths (e.g., support for disadvantaged children) and therefore the language could reasonably be expected to have some similarity. As the data showed that the proportion of similarity was on average half of every report, this was further investigated, as this would be startling news to most school leaders.

Breakdown of similarity percentages

Turnitin describes how much of each report is duplicated from each source it is matched to. The maximum single source was 63% duplication, and the minimum single source 8%, with the spread clustered tightly around the 12-16% mark, with the remaining spanning 22% to 63% (Appendix 4). This implies that there is a core 'bank' of terminology encompassing roughly 15% of a report that is common or possibly mandatory across all reports, and that a small number of reports are almost two third direct duplications of one other report.

This finding led to several suppositions around RQ3 trends in language used in Ofsted reports. With the percentage similarity being high, and the single sources each being substantial, it was estimated that there must be repetition of strings or phrases, some of which probably correlate to mandatory reporting elements from the inspection framework or guidance. As stop words had been applied, titles, dates and structural strings had been eliminated, so these core strings should be within the body of the text rather than headings or footnote patterns. Both Turnitin and NVivo could be set to ‘sentence level’ scrutiny, rather than ‘word level’ - this was applied to identify the strings or phrases (NVivo) and to locate their origin (Turnitin).

Single sources

Those reports with the highest single source data were collated, focussing on those with above the Turnitin threshold for plagiarism (22%). Of the 133 who have a single source of 22% or more, the associated variables were checked, and neither the prior grading, faith designation, HMI or OI author, female or male author, team size or section 5 or section 8 designation showed patterns that were proportionally statistically dissimilar to the corpus. This implies that the duplication patterns are not the consequence of external variables relating to inspector identity or different inspection protocols, but a pattern across the inspectorate.

When geographical trends were analysed, it became clear that a move to regional delivery of inspection had begun, as inspections in roughly the same areas tended to be led by the same lead inspectors. The name of the lead in the highest similarity reports were collated and redacted, to see if there were individual inspector trends, and the name on the report and corresponding lead of the highest single source were added to the spreadsheet. This showed that for the highest similarity reports, some inspectors were using the same language in multiple reports (citing themselves) and some of these duplications were across different inspectors, indicating group language assimilation.

Duplication patterns

These patterns of similarity were investigated and fell into three main styles of duplication. Some were citing themselves at length, duplicating large sections, for example, most of the bullets of the behaviour section or the leadership section would be identical. This was regular in around ten prolific individual inspectors and would periodically appear for other inspectors over the year. This would suggest it is a stylistic choice that some rely on, and which others adopt at certain points. There was no pattern of which reports would show this style of writing, for example, it was seen in each grade awarded, and across all geographical locations, genders, and team sizes.

A second pattern was in the use of very similar sections, usually a bullet point or paragraph in length. These would, in combination throughout the report, add up to a larger percentage duplication overall. For example, a report might include three bullets on page 2, four on page

5 and two bullets each on pages 1, 3, 4 and 6 – which were all duplicated from one other report published by the same inspector, which together make up a self-citation total of 25% of the full report from the same single source. It would be almost impossible to detect that the reports were linked without the software, as the duplications are scattered throughout. These reports also had no contextual trends (gender, location, etc.) but did often have a high overall similarity percentage, as several other sources were also cited in the same report. For example, 25% self-cited from one source, 20% self-cited from another and 18% from a third other report, combined to a total similarity percentage of 63% similar overall.

A third pattern was the use of ‘template writing’ techniques where paragraphs are composed of duplicated sentences interspersed with bespoke writing, or where a stem is copied and given a new ‘leaf’ or ending. The idea, structure, syntax, and meaning are duplicated and slightly amended. This heavily references government documentation, policy writing and formal documentation across the civil service and English regulatory systems (Bazerman and Prior, 2004). It is found in the highest total similarity reports, those with the highest single source and is the most common form of duplication across all reports in the corpus.

Exemplar duplications

The following examples describe the levels and styles of duplication, as the written exemplars best demonstrate the context, content and patterns of duplication found more generally across the corpus. Here, the three styles are exemplified from the list of the inspectors with the highest proportions of self-duplications.

Inspector A

Highest single self-citations in individual reports (63%, 63%, 38%, 34%, 29%) Self-citations are from other published reports by the same inspector within a calendar year. Self-cites whole sections, in some cases almost entire pages. Each ‘borrowed’ segment is more than a paragraph with only single word difference. (Duplicated text, showing the words of difference in bold below)

> *Under the strong* ***and experienced*** *leadership of the headteacher, an ethos of high expectations has been created. Leaders are doggedly determined to eradicate anything that is second best. Leaders' continuing ambitions to eliminate underperformance and to provide the best teaching, outcomes and experiences for all pupils is translated into action, and this means the school is continuing to improve.*

Inspector B

Highest single self-citations from other reports (37%, 31%, 31%, 30%, 30%, 28%, 23%, 22%) Self cites bullet points and paragraphs from several of their other reports. This list reflects the largest single sources, but many of this inspector’s reports are 60%+, made up of roughly 20% duplicated from multiple other recently written reports. This would imply a dominant preferred writing style and syntax choice. (Exemplar, with difference in bold below)

> ***However, these arrangements*** *are very recent* ***and*** *there has not been enough time for the full impact of* ***new*** *leaders’ work to be evident. There are inconsistencies in the quality of teaching across* ***different year*** *groups. This means not enough pupils are making fast enough progress.*

Inspector C

Highest single self-citations in individual reports (40%, 36%, 30%, 27%). Duplicates sentences and bullets periodically, interspersed with bespoke writing. Duplications scatter across the report, implying the use of a ‘bank’ of preferred phrase refined for different schools. A substantial number of other inspectors within the corpus duplicate this inspector at more than 20%.

> *The headteacher* ***and executive headteacher*** *have a passion and determination to secure the best for every pupil.* ***They have*** *established a shared vision of high aspiration and excellence. This is* ***based on a deep*** *understanding of effective* ***leadership****, teaching, and* ***learning****, combined with a thorough understanding of the needs of pupils and families in the local community.*

4.3 Initial theories and refinements

These case studies exemplify common approaches found across the corpus, which are not explicitly banned by the inspectorate. The reporting process includes sending the draft report to an internal Ofsted quality assurance checker, who monitors the report for clarity, errors and reading age amongst other checks. It would be interesting to note which quality assurance person read these highest duplication reports, and what amends had been required through the internal Ofsted quality assurance process, in case the locus of power over the similarity of language choices sits with a specific person or group at the reviewer level, but this would have to be investigated internally by Ofsted as this information is not publicly available.

Case study rejection

Public data shows that there are some schools who statistically have very similar examination outcomes, cohort contexts and staffing or pedagogical structures – these are published on various places such as the government 'get information about schools' webpages or databases such as NCER (National Consortium for Examination Results) or the schools financial benchmarking pages. Schools are required to publish outcomes data on their own websites for parents, alongside their curriculum maps, so it is possible to see some of the potential contextual similarities between similar reports. In this project, the language was tracked quantitatively, rather than undertaking a case study of two very similar reports, to preserve the ethical decision not to focus on individual schools or individual inspectors. To ascertain whether similarities are due to similar schools (e.g., the same cohort needs, same pedagogy, same leadership style and choices) a full interrogation into several schools with highly similar reports would need to be undertaken – this workload is outside of the scope of this review and would be less focussed on the terminology and impact of the reports written as a whole. As such, the line of enquiry was abandoned.

Structure of the report

Those reports who have the highest proportion of similarity show duplicated content throughout the report, in every mandatory section and with no structural trends other than the area for improvement being more unique. For example, the behaviour section does not have more duplication than outcomes. There is a slight bias towards the front-page summaries being more similar than the descriptive content overall, which is likely due to the need to use condensed language to fit into the required single-page format than any bias towards linguistic choices.

For almost all reports the most unique section was the ‘area for improvement’ section, where a description of what to improve is given, and required actions on how to improve. When these sections were separated from the full reports, far more of the area for improvement than any other section of the reports were rated as 0% similar. Discourse analysis techniques would suggest that this implies more power is located within this section. Following this line of enquiry, the ‘area for improvement’ sections were separated from the full reports, and run through Turnitin again as a separate corpus, and against all associated variables, such as size of team, lead inspector, etc.

4.4 Findings: Section 8 as a sub-set

To explore each variable, one element was identified and both Turnitin and NVivo processes applied. This meant that each variable could be explored and compared to the full corpus to ascertain trends or patterns of difference. One of the trends investigated was the comparative similarity of reports from inspections that had begun as full section 5 inspections compared to those reports from inspections that began as shorter section 8 inspections. The assumption by headteachers was that those visits who began as short inspections, where inspectors are not required to write a full report and did not originally plan to complete all the associated activities of a full inspection, would result in reports that relied more heavily on stock phrases and less descriptive language. This was suspected to be due to the need to rapidly add inspection activities when the inspection converted to ‘full’, leaving less time to capture the uniqueness of the setting, thus making the report more generic. This directly applied to possible patterns for RQ1 - uniqueness of reports and could potentially have been a possible answer to RQ3 - language trends within reports.

The section 8 structure was a specific tool designed by the inspectorate to influence governance and improvement, relatively new within the inspection process, and unique to England during this period – so these reports were given additional scrutiny in case themes were found relating to the initial literature review or the wider RQ4 relating to influencing factors on language choices in reports. Within the final corpus, 641 inspections were labelled ‘section 8 deemed section 5’, information included on the ’School Details’ page of the report. The statement referencing the status of the inspection is: *“This inspection was carried out under section 8 of the Education Act 2005. The inspection was also deemed a section 5 inspection under the same Act”*. This complete phrase was added to the ‘stop’ list and so is not included within similarity percentages.

Section 8 deemed section 5 inspections account for between a quarter and a third of all inspections that year (National Audit Office, May 2018). By the next framework update, this process of switching to a full inspection and full report was replaced by a letter advising that a full inspection would occur within 30 months (Ofsted Handbook, 2019). This gave weaker schools a chance to fix issues before re-inspection and indirectly increased significance to the ‘areas for improvement’ section of the letter.

	Section 8 conversions
Led by HMI / Led by an OI	70% / 30%
Led by a Female / Male inspector	54% / 46%
Dropped overall grade	56%
Stayed the same grade	22%
Increased overall grade	22%
Similarity full reports: Section 5 / Section 8	56.4% / 54.3%
Similarity Area for Improvement: Section 5 / Section 8	19.6% / 22.9%

Figure 3 Table showing the proportionate difference in similarity and contextual data for different variables of author and inspection designation

Relatively few non-HMI lead inspectors were qualified to lead section 8 inspections, with the majority being led by HMI, accounting for the skew in HMI lead status. More inspections where the section 8 converted into a full inspection ended with a lower grade. This bias towards a change of grade would suggest these reports should have a more unique description of the school, supporting why the grade has changed compared to a report that describes provision as continuing at the same level.

The full reports for section 5 and section 8 similarity profiles were almost identical. Initial suppositions that the section 8 deemed reports would be more unique was not found to be the case. Although there were some patterns within those rising and dropping grades (see chapter 4.4 ‘changes in grade’) these were not correlated to designation as initially section 8 or section 5.

NVivo investigation into section 8

The NVivo word frequency query was run against the ‘deemed’ inspections as a set, to see if there were any patterns within that set that could answer RQ3 regarding trends in the language of reports - and the list of frequent words that was generated compared to the word frequency of the full corpus. In Appendix 12 the changes in frequent words within the deemed set compared to the frequent words patterns from the full corpus is grouped into statistically significant movement.

For example, in the full corpus, the word ‘good’ is fifth most frequent, but for ‘deemed’ inspections, it is 12th generating a score of -7 (a drop of 7 places). Comparing these lists, the language of the body of the reports varies little between section 8 and section 5 inspections.

Using the iterative process, and applying professional knowledge, words were grouped by theme. The increase in terminology referencing outcomes (outstanding, excellent, assessment, achieve, enough, records, etc.) in section 8 reports reflects the inspector’s need to validate changes in judgement, and terminology referencing the journey from the last inspection (since, time, last, previous) are reduced. There is a reduction of emphasis on ethos (values, caring, behaviour) alongside an increase in emphasis on achievement, which reflects an implicit reference to concrete outcomes evidence supporting a change in grade, something that is more common in section 8 inspection reports, as the main reason for a section 8 to be deemed a section 5 is that a change in grade is probable.

Area for improvement in section 8

For the ‘areas for improvement’ sections in those reports, the same procedure was completed in Appendix 13. The rankings were similar for the top 10 most frequent words, with trends showing core school activity (teaching, learning, progress) similar across both sets. However, for this corpus, the movement of terms was more pronounced. A decrease in the term ‘academy’ (-70) in the larger data is most likely a sampling issue, with fewer academies in the ‘deemed’ set. This is because academisation at this time is relatively new, and academies are required to have a full section 5 for their first inspection within 2 years of opening as an academy, as they are considered a ‘new’ school without a prior grade. Therefore, very few schools in the set will have been academies long enough to be allocated a section 8 to check sustaining of a previous grading (at least 3 years).

Advice relating to re-enforcing areas recognised as strengths (deepen -30, extend -21, strengthen -15) is more prominent in the full corpus compared to section 8, implying that reports for section 8 schools reflect those that have made recent changes which require more time to become established, and the terminology reflects this (secure +20, hold +31, build +20). There is an increase in terms relating to external regulation and monitoring (requirements +27, targets +20, tracking +18, performing +23, measurable +28) which references an increase in directives specifically citing a required focus on quantitative improvement. If this translates into school improvement activity, it could indicate a sharp increase in focus on outcomes and monitoring activity for those schools who have previously been good or better, who were allocated a short section 8 inspection that converted into a full inspection. This aligns with a wider narrative of improvement linked to leadership influence (rigorously, systems, review, responsible, account, policy) and efficiency (funding, premium, additional, effectiveness)

Specific curriculum elements are less prominent in the areas for improvement in section 8 deemed section 5 reports (reasoning -23, grammar -14, handwriting -6, phonics, -7, reading -8, punctuation -9, English -10) reflecting a more general directive rather than specific target given. If this translates into school improvement activity, we should see those schools having the ability to shape their own improvement activity whereas schools sustaining outcomes having to adhere to more detailed and specific improvement targets set by the inspectorate.

Power within section 8 reports

The least similar deemed report is an inadequate report written by an HMI, the most similar a very large ‘requires improvement’ school report written by an OI. This pattern is generally repeated across the corpus, which suggests that HMI, who are used in more complex and difficult cases, have more scope and freedom to describe the context and give bespoke recommendations, whereas OI are more bound by the restrictions of the quality assurance process to refine their language choices into a single, common ‘Ofsted voice’. This is even more pronounced when leading one of the more contentious section 8 inspections and gives the inspector the ‘mantle of authority’ by using more recognisable, inspectorate vocabulary. When an inspection begins as a section 8, the school are not expecting a grade change. Once it is ‘deemed’ this signifies a change is likely, more likely than for an inspection beginning as a section 5, which raises the pressure for the school. The subsequent report then has a much higher level of scrutiny depending on whether the grade did change (validating the decision to deem it section 5) or to persuade the school that a change was not found to be the case, even given the initial decision to deem section 5. In these reports, the inspector wearing ‘the mantle’ of the inspectorate, using language bespoke to the inspectorate, recognisably part of the larger ‘machine’ supports the decisions made by the individuals on site as representative of the overarching system. Here, that is exemplified by fewer specific descriptions and terms, and more generic language referencing systems of regulation and monitoring.

This directly exemplifies the descriptions from Power (2003), Bloom (2017), and Biesta (2013), of the inspector as ‘tool of the state machine’. The findings from this sub-set informed the rest of the investigations, supporting the decision to focus on the ‘area for improvement’ section of the report and giving some direction as to the categories of terminology that might lead to identifiable trends.

Summary section 8

This investigation indicates that the reports from section 8 deemed section 5 inspections follow almost identical patterns to the full section 5 reports, and that leaders' assumptions of reduction in language and bespoke content is not upheld. The influence of changing grades had more impact on difference and language choice than whether the inspection began as section 5 or section 8, and there were only marginal differences in language choice, with areas for improvement slightly more generic and less focussed on curriculum areas. This makes these reports more difficult to track against concrete pupil outcomes for RQ5 impact of reports and language on outcomes. In Chapter 5 I discuss some potential indirect consequences of this finding.

4.5 Patterns in those reports where there are changes in grade

To further investigate RQ3 trends in the language used in reports, a subset of schools who had improved their grade or decreased their overall rating were qualitatively analysed with a focus on language and its power and implications. The supposition was that those schools who had been 'demoted' in grading might share common linguistic terms that could then be used for the benefit of other schools to avoid similar consequences. Equally, if there were any linguistic similarities between those schools who had been successful, then others could see the topics or common descriptions and subsequently also be successful.

The 'area for improvement' sections were scrutinised, as that section of the report describes the actions required to improve out of a lower grading, and which, if Ofsted's reporting system does lead to improved schools, should represent the most powerful language to direct improvement.

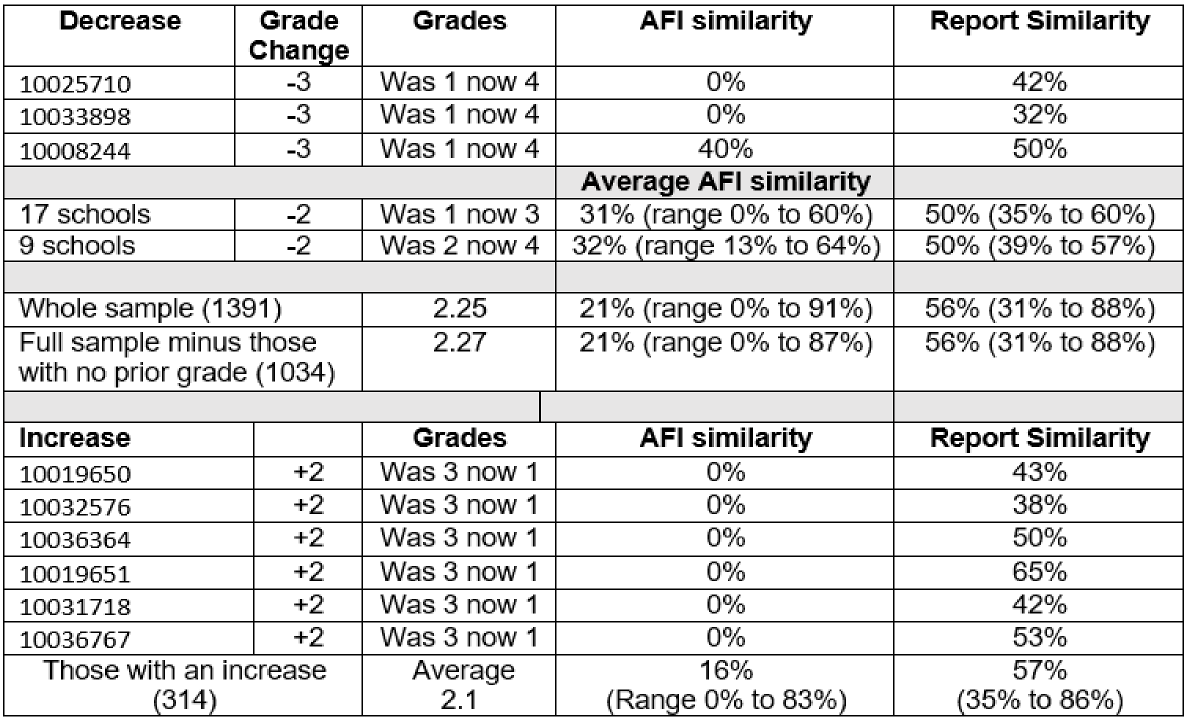
Schools who had dropped more than one grade were 29 out of the 1391 total (2%), or 29 out of 1034 (3%) if we discount those with no previous grade as they are new schools. For these schools with the biggest grade change, the 'areas for improvement' were collated and averages taken. Those changing more than one grade had areas for improvement on average more unique compared to the corpus. Those that decreased a grade had full reports that were marginally more unique than the corpus overall (Figure 5). This implies that schools who receive a grade change following inspection are given improvement advice that is more specifically tailored to their unique circumstances. However, the main body similarity percentage is the same as the corpus, suggesting that the section descriptions remain generic, and the assumption that reports that change grades would need to exemplify those changes with unique text in the body of the report is incorrect.

This suggests that the areas for improvement are more critical to improvement and therefore more bespoke and therefore more helpful to the school than the body of the report.

Changed grade areas for improvement

Due to this, the ‘area for improvement’ were investigated using NVivo for language trends. This was to see if there were any linguistic patterns or words that were particularly associated with these areas for improvement, in case they carried more power, influence or ability to drive school improvement than other areas.

The top 10 most frequent terms driving required actions are very similar to the full corpus of ‘areas for improvement’ in every subset and variable. This is possibly because the most frequent words relate to the requirement to describe core actions in terms relating to the judgement areas. For example, ‘Teaching, Learning and Assessment’ is one of the core judgement areas, and all three are in the top ten of all ‘areas for improvement’ search lists. Equally, ‘improve’, ‘pupils’ and ‘school’ are the top three terms, used in almost every ‘area for improvement’. Therefore, these most frequent terms represent the automatic inclusions rather than the interesting differences. This implies common actors and actions, but more unique combinations of terms, and that there are no word-level trends particularly associated with substantial grade changes.



Decrease	Grade Change	Grades	AFI similarity	Report Similarity
10025710	-3	Was 1 now 4	0%	42%
10033898	-3	Was 1 now 4	0%	32%
10008244	-3	Was 1 now 4	40%	50%
			Average AFI similarity	
17 schools	-2	Was 1 now 3	31% (range 0% to 60%)	50% (35% to 60%)
9 schools	-2	Was 2 now 4	32% (range 13% to 64%)	50% (39% to 57%)
Whole sample (1391)		2.25	21% (range 0% to 91%)	56% (31% to 88%)
Full sample minus those with no prior grade (1034)		2.27	21% (range 0% to 87%)	56% (31% to 88%)
Increase		**Grades**	**AFI similarity**	**Report Similarity**
10019650	+2	Was 3 now 1	0%	43%
10032576	+2	Was 3 now 1	0%	38%
10036364	+2	Was 3 now 1	0%	50%
10019651	+2	Was 3 now 1	0%	65%
10031718	+2	Was 3 now 1	0%	42%
10036767	+2	Was 3 now 1	0%	53%
Those with an increase (314)		Average 2.1	16% (Range 0% to 83%)	57% (35% to 86%)

Figure 4 Table showing those with the largest changes in grade.

Patterns in those reports where grades have decreased

Those schools who decreased more than one grade were investigated using word frequency to see if there were linguistic trends or spikes unique to this group. From the top 200 most frequent words there are some interesting additions for those schools that drop grades compared to the full set of areas for improvement. For example, policy, level, target, opportunities, and funding all appear in the top 200 most frequent words for dropped schools, when they are not in the top 200 of all areas for improvement. This implies a considerable shift towards changing core approaches (policies) and a focus on quantitative measures (target, level) for dropped schools is required. These are actions that could potentially change a schools' overall culture or ethos. The new appearance of 'funding' implies a future monitoring of finances in a way that could influence a school's spending patterns to prioritise short-term resolutions (i.e., by the next inspection, which for these schools is within 2 years) over longer term financial planning or sustainability.

Words relating to the process, speed and monitoring of change (evaluation, precisely, rapidly, sufficiently, rigorously) are newly prevalent in the dropped grades set. The emphasis on monitoring is important, and the requirement to not only evaluate, but to assign accountability (responsibility, roles) also appears in this set but not the full corpus. These schools are required to not only extend their quality assurance practices, but to hold staff accountable in a way that can be overtly monitored and evaluated by the next inspection. This could have an impact on staffing, workload and on performance management during a period of extreme change and potential instability.

Words relating to teaching (planning, marking, consistent, challenge, present, duties) and new words relating to learning (extend, accelerate, performance, confidence, attaining) also appear, demonstrating that required actions are tied to core teacher activity. These words appear in the teacher standards document active at the time (Teachers' standards 2013), which implies that staff in these settings may not be fulfilling the teacher standards, and that the issues with the school's performance lies with the main staff body. This is reflected in the language where leadership is implied but not cited e.g., 'ensure that teachers…' Which clouds accountability between the leaders who should be 'ensuring' or the teachers who have not done whatever is described. The placement of power is difficult to ascertain, and the language of these 'areas for improvement' related to teaching and learning places the power for change either with teaching staff or their direct leaders. Children are described as passive recipients of the outcomes of these changes, or simply the conduit of the intervention.

> *Improve the quality of teaching, learning and assessment by making sure that teachers:*
>
> *- Set work for the most able pupils that is sufficiently challenging,*
>
> *- Make better use of resources, including the learning environment and staff, to support pupils' learning, particularly to develop pupils' information and technology skills,*
>
> *- Build effectively upon pupils' early phonic skills so that they are able to spell and read unfamiliar words correctly.*

From inspection #10025598 showing a 'stem' of an overarching outcome for teaching, with a very specific action in the 'leaf' that would dictate a concrete outcomes-focussed pedagogy to be able to accomplish successfully within the timescale before re-inspection (18 months)

> *…read unfamiliar words correctly.*
>
> *Improve the quality of teaching and assessment by ensuring that:*
>
> *- All classes are taught by high-quality, permanent members of staff who will provide continuity and stability to pupils' learning.*

From inspection # 10003046 showing an overarching stem for teaching with a very specific action of employing permanent teaching staff. This is a financial and strategic decision outside the scope of regulation yet was included in the schools 'required actions' that formed the basis of its re-inspection. It implies a causative impact between employing permanent members of staff and pupil's learning being 'stable' – children are the passive recipients of this action.

Some core groups appear as frequent terms in this set, (boys, middle, nursery, reception, higher, special, abilities) whereas they are not frequent features of the wider 'areas for improvement'. This is an indication of very directed and specific required actions linked to measurable groups of pupils for these schools. This would require leaders to have an enforced priority on quantitative outcomes for these pupils and thereby influence school development plans to include these groups as critical, even if in current or future cohorts they represent a different population of abilities or needs. Whether these 'groups' are consistent over time for these schools or if they are representative of significant populations is not clarified within the reports, and the old quantitative measures for Ofsted terminology (e.g., 'most' 'a substantial proportion' etc, Subsidiary guidance #110166, 2012) have been removed from formal inspection documents at this point, so the relative proportions of these groups compared to the student body as a whole are unknown.

Subject specific terminology (spelling, science, outdoor) are new additions, or much higher in the rank order (phonics +26, curriculum +28, assessment +35), whereas safeguarding drops 96 places, parents -107, and community -73. This echoes the pattern of prioritising outcomes over culture or ethos and requires schools that have dropped a grade to skew their improvement priorities towards specific, measurable curriculum elements. As these schools are mostly due re-inspection after only one external data point (e.g., SATS exams) this emphasis on these curriculum elements or groups of children will be considerable.

For example, the 'areas for improvement' from this dropped grade inspection (#10025710) in December 2016, published January 2017,

> *Improve outcomes, particularly in writing and for boys, by:*
>
> - *developing better writing skills*
> - *ensuring that the outdoor provision in Reception promotes early writing,*
> - *adapting the curriculum so that boys are much better engaged in their learning.*

relates to outcomes from the children in the academic school years starting September 2016. Therefore, in order to 'improve writing outcomes for boys by promoting early writing in reception's outdoor provision', leaders would have to significantly accelerate the progress and attainment of current reception children (on whom the judgement was made) by the end of this academic year (less than two terms remaining) in order to be verified by national data and then influence the writing of the incoming reception year group, a cohort new to school, with as yet unknown writing levels, who would potentially not have completed a full year by the point of next inspection. This influences the school's strategy towards short-term, interventionist approaches, as the only strategies that could secure demonstrable improvement in pupil data in the limited time frame available. This could pressurise academic achievement over the pastoral or wider needs of this cohort of 5-year-olds to secure the school's reputation and future.

Iterative investigation of RQ 2 and 3

At this point, the analysis was moving from RQ3 investigating the trends in language and started to look at RQ2 whether the language in reports leads to subsequent improvements. It is this critical impact of the area for improvement on school strategies and priorities from my own practice that initially led to the creation of this project. This area for improvement does not account for short-term cohort changes, for example, the incoming reception cohort may arrive with stronger writing skills, and so may show limited progress within external measures, or higher attainment may look like the result of changes made but could be down to cohort variance, and vice versa for lower scores. Because of this, an increased importance is placed on teaching strategies and overt, observable (or easily evidenced) short-term intervention so that the school can demonstrate a response to this area for improvement outside of external data, which has a high risk of not being able to evidence impact in the time available.

Stem and leaf

In this case, the ‘stem’ of the area for improvement “Improve outcomes…” covers the required endpoint, and the ‘leaf’ “…by ensuring the outdoor provision in reception promotes early writing” refers to the actions required. This pattern is often replicated in the area for improvement of schools who dropped grades, with the desired impact being stated, with concrete, monitorable actions following.

If the school chooses not to change writing approaches in the outdoor provision, then leaders will be held to account if outcomes in writing do not improve. However, if the outcomes for writing do improve, then it is unlikely that the inspectors could tackle the school for not changing outdoor provision, as it would be dictating pedagogy, which is specifically cited as not the purpose of inspection within the handbook (Ofsted handbook, 2018). It would be risky for a leader to not have an action on their development plan to change outdoor writing strategies, even though this would then be the inspectorate dictating a specific pedagogical approach. This appears to be also prevalent in ‘deemed’ inspections.

The stem and leaf design requires leaders to read each part as a separate (albeit linked) target requiring action. This ‘area for improvement’ equally gives a generic target of ‘Improve outcomes, particularly in writing…by developing better writing skills”. It would be almost impossible to say how the inspector required writing to be improved from reading this portion of the area for improvement, and which skills, or which outcomes they are referring to. This could be in-year progress, end of Key Stage attainment, writing across the curriculum or only in English, there are multiple interpretations of this directive. It is in fact so generic, that leaders could successfully improve, for example, handwriting, when the weakness was in writing at length in subjects other than English. However, the inspector who returns to the school to re-inspect and who will use these areas for improvement as one of their trails, will also only have the generic statement as their starting point, and only the interpretations of the leadership team to base the evaluation of the school’s successful (or not) addressing of this area for improvement.

Similarly, the stem includes boys as a specific group. In this case, it would be easy for leaders to prioritise boys' achievement, putting short-term efforts and actions in place for boys alone. It would also be likely that leaders would focus on boys writing, as the two are linguistically linked here as if the two were congruent. However, the full reading of the stem requires improvement for all pupils, and for all pupils' writing and only the boys are particularly weak within the overall group. The subsequent 'leaf' descriptors specify that it is the writing skills of all pupils, the early writing outdoors in Reception and the fact that boys are less engaged that are the problem areas. Taken without the 'stem' the area for improvement would have much less of a focus on boys writing and could lead to a substantially different interpretation of the required actions. The placement of the terms here could skew leaders' interpretations of the area for improvement, meaning that the most powerful words, which influence improvement and leadership actions the most, if they are written in an ambiguous way, could lose or misplace their impact.

Those new to headship, having their first inspection, or un-supported by more experienced readers of 'areas for improvement' could easily mis-interpret actions written in this way. In my practice, during monitoring visits of special measures schools, an HMI would often have to sit with the head and un-pick not only the language and content but would re-visit the discussions during inspection that clarified where these targets had been informed by the observations, book scrutiny and discussions during inspection. Often, heads forget key discussion points due to the pressure of the inspection regime, and the prospect of dropping grade. As an inspector, I would invite a member of the leadership team to accompany the head into meetings to capture copious notes, so that the head could focus on following the connections from evidence collected to the area for improvement. In practice, the wording of areas for improvement were often changed during the quality assurance process to ensure word limits were adhered to, educational terms were explained, etc. In recent years, during final feedback, it is common to hear the lead inspector say something like '*the wording may be changed, but the essence of the areas for improvement will remain the same*'.

Summary decreased grades

Although few trends in terminology were found, the importance of word selection in areas for improvement are highlighted. Where inspectors use language that is in any way ambiguous or lacks description, it is possible for leaders to mis-interpret or mis-direct improvement activity. This will influence the evaluation of RQ2 'Does the language of reports lead to subsequent improvements in outcomes', as reports with ambiguous language may not lead to the improvements required. Investigation into ambiguous areas for improvement, or those who, when responding to feedback forms following inspection, indicated that 'the written report did not match the verbal feedback' question would be an interesting addition to the knowledge base. However, this information is not publicly available, and so was not included in this project.

Patterns in those reports where grades have increased

As there had been some indication that where reports had increased grades, patterns of language may be different, these were looked at as a sub-set. This investigated RQ3 looking for trends in the language, to see if those schools who are doing well had specific terms that may reflect actions or issues that could be used to support other schools to do equally well.

314 schools increased their grade, and the full reports of these schools were within 1% of the similarity of the corpus, suggesting the body of the report reflects the same language replication patterns as those schools who remain at the same grade. The 'area for improvement' for these schools were more unique than the full corpus, with 163 out of 314 (52%) at 0% similar, (compared to 42% for the full corpus), although a small number of very similar areas for improvement skewed the overall average for this sub-set. The few very similar areas for improvement (as high as 83% replicated content) that skewed the area for improvement average for increased grade reports replicated whole sections of similar areas for improvement, made only partially bespoke with signifiers.

This re-enforced the assumption that there are common core improvement activities or strategies that will move a school from one judgement grade to a higher one, or that are required for outstanding schools to sustain their top judgement. These were investigated to see if they reflected any educational trends or fashions, as those things that are suggested as options for schools to sustain an outstanding grade are likely to include staying up to date with the latest best practice This could be a way for schools not yet inspected to identify the core difference between 'good' and 'outstanding' and would inform RQ2 'Does the language of reports lead to subsequent improvements' and could inform RQ4 'Are there any influencing factors on the language used?'.

The 'area for improvement' for schools increasing their grading were 15% more unique than those with reduced grades, and even more unique for those schools who uplifted two grades. As these are a statistically small sub-group, it is difficult to interpret consistent significance from this, although it would make logical sense that it would take much more detailed language to explain a significant rise, reflecting a broader and more unique evidence base to explain the increase.

An NVivo word frequency search of these reports shows that key terms from the Teachers' Standards are frequently used, including 'consistent', 'plan' and 'progress', which appear in similar frequencies to the full corpus. Language that pertains to established practices in school is also frequent (e.g., 'continue', 'sharpen', 'more', 'fully'), which is to be expected from already successful schools. The relative similarity between the terminology in the area for improvement of those schools whose grade improved and areas for improvement as a whole, implies that for those schools who are successful and improving, the advice given pertains to generic school areas and not to any new or different practices.

The uniqueness of the 'areas for improvement' sections seems to come from the inclusion of establishing signifiers such as 'continue to' or 'further develop' rather than any new or bespoke content. This could indicate that the opportunities for positive language in areas for improvement are limited, because the section is required to only represent direction and actions. Language such as 'continue to' and 'consistently' imply that good aspects are in place and so these terms occur more frequently in reports on good or outstanding schools.

The example below is taken from inspection #10025181, which was found to be 73% similar to other reports in the corpus. Here, whole phrases are replicated, with only establishing signifiers changed, (e.g., 'even stronger', 'focus more'):

Improve teaching and learning by ensuring that:

- *Teachers **consistently** check pupils' understanding during lessons and use this to shape learning effectively,*
- *Pupils are set **consistently** challenging tasks to extend their learning and progress, **especially the most able** pupils.*

The example below is taken from inspection #10024518, which was found to be 83% similar.

> ***Continue to improve** the quality of teaching to ensure that pupils make **the best possible** progress by:*
>
> - *making sure that teachers set the right level of challenge to enable pupils to make **even stronger** progress…*
> - *…sharpening leaders' monitoring of teaching so they focus more on how effectively **groups of** pupils learn and make progress in lessons…*

These phrases in particular are used extensively in a wide range of slightly amended format, all pertaining to a required improvement for feedback and challenge.

Quality Assurance

One aspect of the Ofsted quality assurance process is to check that the areas for improvement given in the report do not undermine or contradict the overall judgement. For example, in a school that has been raised out of a 'Requires Improvement' or lower grade, any area for improvement that might indicate a substantial area that still requires improvement would contradict the overall increase in grade. Therefore, these areas for improvement sections will have been scrutinised for their scale and scope, refining actions into smaller or specific sub-sets in order to limit their impact on the overall judgement. For example, "improve *boys* writing *in early years*" rather than 'improve writing'. This could account for the high level of establishing signifiers used, to minimise the scope of the area for improvement, and imply generally positive practice.

Summary increased gradings

Analysis of those reports from schools that improved their grade shows that there is a difference in the specificity or ambiguity of language used in the area for improvement sections within this subset. More unique, yet generalised actions relating to leaders rather than provision are found in reports on weaker schools, and more specific and directed actions are found in reports in those schools that are already good. Where language is replicated, these are commonly references to the Ofsted handbook bullet points for 'good' evaluations, with establishing signifiers used to minimise the scale of requests to improve. Use of terms such as 'consistently' and 'especially for…', for example, will influence school decisions on the focus of improvement activity and are therefore powerful within the areas for improvement section.

4.6 Inadequate schools

A full scrutiny of reports on schools graded 'inadequate' within the corpus was undertaken, as existing literature indicates that inspection had the most impact on schools with this grading across Europe (Ehrens, 2014). As 'inadequate' is the lowest possible grading in the Ofsted framework, these schools will have required the greatest degree of improvement of all those within the corpus. Therefore, reports on those schools graded inadequate should be the richest source of terminology that inspectors expect to lead to improvements (RQ2). It is also expected, therefore, that terms that are the most 'effective or impactful' will be identified (RQ5). If inspectors are aware (even subconsciously) of terms that will lead to the best improvements, they should be found frequently within the reports that advise the worst performing schools on how to improve.

The corpus reflects the low proportion of schools graded ‘inadequate’ nationally, and quantitative trends related to these schools are statistically insignificant within this overall group. However, as a sub-set, some trends are evident. Of the 15 schools with an overall ‘inadequate’ grading, only four are inadequate in all areas. The rest have some strengths recognised, with higher gradings to specific areas. As such, their areas for improvement should be specific to the areas affected by the lowest judgments as indicated by the Ofsted handbook. The data shows that the areas for improvement of reports on schools graded inadequate are more similar than the areas for improvement across the corpus as a whole. The full reports are much less similar. This is potentially due to Ofsted having to outline to leaders how to improve those things that have been deemed inadequate, and so the language used will closely reflect the wording of the inadequate judgement section of the report in the areas for improvement section. For example, where the inadequate judgement reads, “The range of subjects is narrow and does not prepare pupils for the opportunities, responsibilities and experiences of life in modern Britain” (Ofsted Handbook, 2018), the corresponding area for improvement is likely to be written as follows: “Improve the breadth and range of subjects to prepare pupils more adequately for the opportunities, responsibilities and experiences of life in modern Britain.” This requirement for consistency of terminology both limits word choice and makes the areas for improvement section less unique, as reports will replicate this exact wording.

As indicated above, the difference in the level of similarity from full reports could be due to the higher word count allowed for inadequate reports, to enable inspectors to clarify exactly why a school’s performance is being deemed inadequate. Inspectors are required to provide clear evidence for each inadequate factor, as this is a critical descriptor not only for school leaders but also for governors, parents, and the wider audience of a report. For example, report #10033898, which reports an inadequate judgement, has 5,573 words, compared to a ‘good’ school report, such as #10036957, which has only 3,667 words. Word count does vary across the corpus from less than 3000 to almost 6000 and although it cannot be statistically confirmed that reports for schools deemed inadequate are always longer, within this specific corpus, reports for schools deemed inadequate were longer and less similar (Appendix 6).

Areas for improvement Inadequate reports

Within the areas for improvement of reports on schools deemed inadequate, there was a higher frequency of use of terms relating to leadership and management (‘leadership’ +11 places; ‘managing’ +15) compared to areas for improvement overall, which implies a need to increase actions focussed on ownership and accountability for improvement. A decrease in terminology relating to outcomes (‘progress’ -5, ‘quality’ -6, ‘develop’ -16) and pedagogy (‘subjects’ -40, ‘maths’ -26, ‘reading’ -22, ‘writing’ -14) also reflects a vocabulary for this sub-set that focuses on actors rather than impact. This more directive language is often formed by a stem that describes the actor (e.g.), followed by a concrete action that can be easily evidenced. Many areas for improvement start with the stem ‘Leaders and managers should…” or synonyms for actors and ownership of improvement. Although there was increased reference to children (+15) this tended to position them as passive recipients of activity and as a measurable factor (‘assess’ +7, ‘needs’ +13, ‘check’ +30).

The areas for improvement in reports on inadequate schools include a combination of overarching, difficult to evidence statements, and clear, concrete required actions. This shows that although these sections are more bespoke in terms of the language used, the actions required to make these improvements could be more variable than other ‘areas for improvement’ across the whole corpus. Although the improvement section for inadequate schools is almost always significantly longer, the areas they describe seem to be broader and more generic. This pattern seems to indicate that the areas for improvement for inadequate schools heavily utilises the Ofsted handbook terminology to describe inadequate areas, assuming that specific guidance to improve those areas can be found in the associated parts of the handbook and body of the report. As many relate to core school function or standards (teaching, managing, basic provision) the details of ‘what good looks like’ can be found in other documents, and the ‘area for improvement’ is listing inspection focus for returning inspectors.

Issues with generic areas for improvement

As some statements are quite broad and generic, it is difficult to say what evidence a school would need to produce in response to these broad headings in order to improve their grading from ‘inadequate’. For example, in #10023528 ‘provide a rich and broad curriculum’ suggests that the school is currently not providing this but does not specify what is missing from the curriculum they are offering. Whereas in areas for improvement of other reports, what is missing is clearly stated, and what to do is given in concrete terms, these overarching statements leave schools to interpret the required actions. If we take the ‘rich and broad curriculum’ example further, within the full report, no omissions or required additions to the curriculum are indicated in that report. Insufficient progress in reading, writing and mathematics in KS2 is described, while Early Years provision is judged to be ‘good’, and KS1 outcomes are said to be ‘in line with national’. P.E. and computing are celebrated, and the word ‘curriculum’ only appears in the areas for improvement section and on the front page. The front page of the report reads:

> *Although pupils study a wide range of subjects, the quality of work seen in books shows that too few pupils achieve well in the* ***wider curriculum****. Too many teachers have low expectations of what pupils can achieve in subjects such as science and geography.*

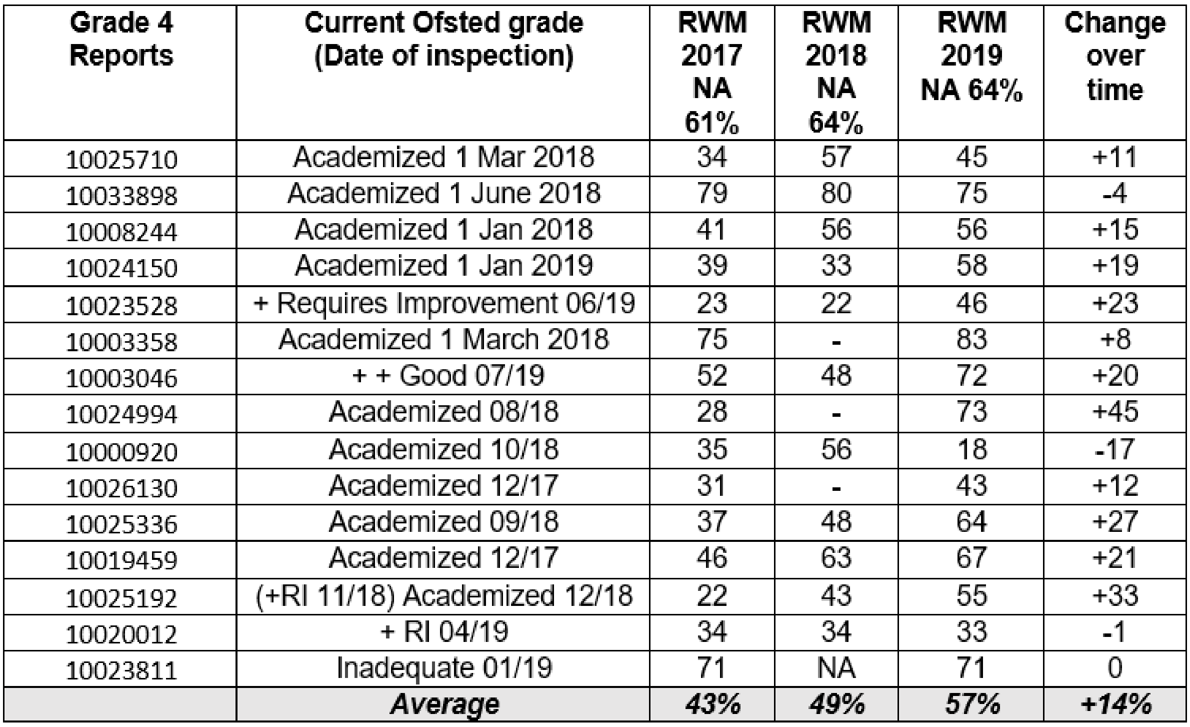
The corresponding ‘area for improvement’ reads:

> *Improve the quality of leadership and management by:*
> *– providing pupils with a rich and* ***broad curriculum***
> *Improve pupils’ outcomes by:*
> *– raising teachers’ awareness of what pupils can achieve in subjects across the* ***curriculum****.*

(Report #10023528)

Here, science and geography are mentioned as insufficient. In the Teaching section of the report, history outcomes are described as weak. In none of the cases is curriculum provision described as not being ‘rich’ or ‘broad’, however, and no omissions are described. Some lack of inclusion of ‘other religions’ is stated, which falls within the inspection framework requirement to promote ‘Fundamental British Values’.

Schools graded higher than inadequate appear to receive more detailed feedback and advice and more concrete or easily evidenced required actions that they will be judged against in future inspections. For schools judged inadequate, the similarity of language may mean that the report and the areas for improvement have less impact. Although the school is required to respond to the ‘areas for improvement’, leaders of schools judged to be inadequate appear to have more freedom in how they choose to respond. Due to this, it is difficult to say if any specific terms are implied to have an impact on school improvement. However, it is possible to evaluate the impact on outcomes in core subjects of this use of more generic advice, as these are a matter of public record. Hence, outcomes in core subjects can be used as a proxy measure for the overall impact of the wording of these reports.



Grade 4 Reports	Current Ofsted grade (Date of inspection)	RWM 2017 NA 61%	RWM 2018 NA 64%	RWM 2019 NA 64%	Change over time
10025710	Academized 1 Mar 2018	34	57	45	+11
10033898	Academized 1 June 2018	79	80	75	-4
10008244	Academized 1 Jan 2018	41	56	56	+15
10024150	Academized 1 Jan 2019	39	33	58	+19
10023528	+ Requires Improvement 06/19	23	22	46	+23
10003358	Academized 1 March 2018	75	-	83	+8
10003046	+ + Good 07/19	52	48	72	+20
10024994	Academized 08/18	28	-	73	+45
10000920	Academized 10/18	35	56	18	-17
10026130	Academized 12/17	31	-	43	+12
10025336	Academized 09/18	37	48	64	+27
10019459	Academized 12/17	46	63	67	+21
10025192	(+RI 11/18) Academized 12/18	22	43	55	+33
10020012	+ RI 04/19	34	34	33	-1
10023811	Inadequate 01/19	71	NA	71	0
	Average	***43%***	***49%***	***57%***	***+14%***

Figure 5 Table showing the combined performance in KS2 reading writing and mathematics statutory assessment tests from 2017-2019 for those schools graded inadequate in 2017. NA indicates the national average scores for those years.

An average change of +14% change in combined Key Stage 2 outcomes can be seen in the 'inadequate' sub-set from 2017 to 2019. In the year following inspection, four of the 15 inadequate schools did not post results, which is normally an indicator that they had so few pupils in examination year groups that the data would have made individual pupils identifiable. (One effect of a school being placed in special measures is a drop in pupil numbers (Wilcox, 1996) which makes comparisons of improvement difficult.) In the year immediately following inspection, 7 out of the 11 schools for which outcomes data is available showed an improvement, a 6% increase overall. Two schools were judged higher than inadequate within 18 months. Whether this is due solely to following the 'areas for improvement', or because of academization is unclear. Eleven out of the fifteen schools judged inadequate became academies in the period following inspection, joining pre-established trusts as 'sponsored' academies.

Of those that did not academize, #10003046 had concrete areas for improvement that related to appointing permanent staff, teacher's standards and safeguarding, and improved its judgement to good within 2 years. Report #10020012 had actions relating to safeguarding, teacher's standards, and risk assessment, and improved to Requires Improvement in two years, despite little change in outcomes as measured by core subject data. Report #10023811, which had some concrete areas for improvement for assessment and some generic requirements for curriculum and expectations, was judged inadequate again on re-inspection two years later, despite no change in outcomes, which remained above national average. These schools are worthy of further investigation as critical outliers within this sample.

Sentiment analysis inadequate reports

Sentiment analysis on the improvement section of the reports in the 'inadequate' sub-set showed a skew towards positive vocabulary. This reflects a syntax that is structured around aspirations (improve, engage, create, support). In only three reports are there direct references to the negative findings (accuracy, efficacy) that underline the judgement.

By contrast, on the front pages of the reports on inadequate schools the language is more negative and gives a much harsher picture of the school's weaknesses. Report #1025192 is an example of this. The opening phrase of the area for improvement reads:

> *As a matter of urgency, ensure that all aspects of safeguarding are effective by:*
> *…updating the safeguarding policy to ensure it reflects current government guidance and the school's own context.*
> *…providing adequate supervision at breaktimes and lunchtimes to keep pupils safe.*

Whereas the areas for improvement were evaluated as a 3, 'moderately positive', in the sentiment analysis, the front page has a very different tone:

> *Safeguarding is inadequate. Too many staff have not received the necessary training to know how to keep pupils safe. Some staff do not understand the safeguarding policies and procedures. Processes to recruit suitable staff are not good enough. Pupils are not supervised properly at informal times.*

The required actions stated in the 'Areas for Improvement' are strong and specific and can be evidenced by concrete outcomes on re-inspection, yet the tone is much softer than expected, and the language refers to policy documents and actor responsibilities rather than an impact on children. In good schools, by contrast, 'areas for improvement' show there is a much stronger bias towards describing the required impact on children. Just as in reports, on schools whose grade improved, where establishing terminology (e.g., continue to…) softens the impact of the directive, here, the qualifiers (e.g., adequate supervision, at breaktime…) appear to soften the tone. There is a considerable difference between 'properly supervise at informal times' and 'provide adequate supervision at breaktimes and lunchtimes to keep pupils safe' in terms of the power of the language. The specificity of the latter provides a focus on times of day and the nature of the existing supervision, i.e., there is supervision, but it is not fully adequate. It is possible that the character limit to fit descriptions onto the front page are drivers of condensing the terminology used, but this pattern replicates across many of the inadequate reports.

Summary: Inadequate reports

For those schools within the corpus graded ‘inadequate’ their areas for improvement were more similar than the corpus as a whole. They had a higher word count overall and within the area for improvement section. Areas to improve were broad and generic, and heavily reliant on the Ofsted handbook terminology. The areas for improvement focussed on actors (leaders, managers, teachers) rather than outcomes for pupils. This group had a strong improvement in outcomes for pupils following inspection, but from a very low starting point.

4.7 Leads and Teams.

In relation to the question of whether there were trends in Ofsted reports (RQ3), the status and gender of the lead inspector were investigated, as similar discourse analysis studies cited gender and seniority of the report author as common influencing variables (Bazerman and Prior 2004). The data in Appendix 5 shows levels of similarity between male- or female-authored reports or between full-time inspectors (HMI) or part time Ofsted inspectors (OI) are negligible overall. Similarity scores suggest that HMI reports are occasionally more unique than OI reports. Given that the average similarity of HMI reports (-2.1 from corpus) is less than the average of OI (+2.3) this would appear to back up that evaluation.

T tests were used to check whether gender and status are significant variables. For both full reports (P= 0.3, n=1391) and areas for improvement only (P= 0.8, n=1391), the T-test for gender found no statistical significance. This eliminated gender as a line of enquiry to further investigate trends in language and factors influencing language (RQ3 and RQ4). Similarly, sizes of teams and similarity profiles were eliminated as not statistically significant as a variable that impacted on the uniqueness of a report or an ‘area for improvement’ section.

As team size and status of lead are linked, with HMI taking the majority of very small and very large inspections, and as the literature suggested that there were influencing factors on reporting that were common for larger teams, this was explored. In terms of authorship, there are marginally more HMI leading larger teams (4 and above) which could reflect the increased complexity of managing a larger team, or a re-inspection, which accounts for a small number of very large teams. The size of the inspection team is calculated on a pro-rata basis, based on the number of pupils on roll. In some cases, an inspection may be repeated or extended to collect additional evidence, called a ‘split’ inspection (e.g., if a school closes for snow mid-inspection). In these cases, if the original team cannot return, any additional inspectors brought in to complete the inspection are included as extra team members on the report. In this corpus, because it focuses on primary schools, who have a relatively smaller number of children on roll, the inspection teams are generally 4 inspectors or fewer. Those with more than 5 inspectors in this corpus reflect split inspections. For split inspections, the original lead tends to be maintained, even if an HMI or second lead is added. The size of the team was hypothesised as having a possible impact on RQ3 trends in language or RQ4 influence on language, in line with the literature review, which suggested that, where there is a single writer, a lone 'voice', patterns are linguistically different to group-written texts (Bazerman and Prior 2004). On a larger team, evidence is collected and analysed by multiple 'voices' and discussed on site, which could influence the language choices used in the final report. This potential indicator of influence on the language and/or content of reports was investigated using NVivo to find trends in language use, terminology similarity or difference in sentiment, and differences in the language choices noted. Although there were some spikes in the data, these correlated with previously discussed variable patterns, such as inadequate schools, rather than presenting as new trends. The T-tests between different team sizes for report or area for improvement similarity, word frequency and sentiment confirmed group size had no discernible significant difference to other variables and confirmed the theory that the 'voice of the inspectorate' had almost completely eliminated the impact of individual authorial variables.

4.8 Area for improvement

As the 'area for improvement' section was emerging as critical to answering questions on uniqueness and trends (RQ1 and RQ3), further analysis was conducted on this subset. The analysis of full reports had shown that there were few authorial trends, very similar word choices, and three main patterns of duplication – block duplication, self-referencing and template writing. However, within the areas for improvement, these patterns were not present and there was more unique content. Therefore, these sections of the report were potentially more impactful. Hence, specific attention was given to the areas for improvement of the reports in the corpus.

Sentiment coding and auto-coding (NVivo) had shown that the trending language and majority of significant outliers were all to be found in the areas for improvement section of reports, indicating that this small section of the larger report carried disproportionately more power and influence when considered from a discourse analytic perspective. The 'areas for improvement' sections were coded as a separate corpus, so that this could be investigated. This new corpus was run through Turnitin and NVivo, and similarity reports and word frequencies were generated, as had been done for the main corpus.

The areas for improvement sections, being shorter and already identified as more bespoke, showed different trends to the full reports. Of the areas for improvement, 584 were shown to be totally unique (i.e., 0% similar), within which the proportion of grades was comparable to the full corpus, with slightly more outstanding and good, and slightly fewer 'requires improvement'. This seems to confirm that those schools who are successful are more likely to have bespoke advice on improvement, whereas those schools that require improvement are likely to receive broad recommendations that are more generic or are highly similar to actions required by other settings. This is an important finding, as it implies that the inspection process requires schools to follow generic advice and actions to achieve a 'good' Ofsted grading, and that once good, schools then require bespoke, unique advice. This finding does match my professional experience, whereby those schools who are less than good have had to work towards broad, generic targets that are, in essence, summaries of the Ofsted inspection criteria for 'good' rather than specific, personalised recommendations based on the inspectors' observations of the school's particular circumstances or context.

Most frequent duplication within 'areas for improvement'

For those 'areas for improvement' who scored highly for similarity percentage on Turnitin (over 70%), with the highest percentage of text that previously exists within other 'area for improvement' sections of the same year, a further analysis was undertaken to see if contextual factors or professional knowledge could explain these spikes in the data. In this section, there are mandatory elements. For example, where there has been inadequate governance or pupil premium funding handling, set phrases exist which must be included:

> 161. Inspectors will recommend an external review if governance is weak. Under 'What the school should do to improve further', inspectors should use the following words in the report: *'An external review of governance should be undertaken in order to assess how this aspect of leadership and management may be improved'.*
>
> Ofsted Handbook, 2017 page 45, section 161

As this phrase includes core vocabulary that cannot be omitted without detriment to frequency searches and trends, this could not be discounted as a 'stop phrase'. This does not skew the data on similarity overall, as the phrases duplicated are identified individually, and the mandatory phrases can be identified as specific trends, separate from inspector-chosen terminology.

The nature of 'area for improvement' sections is that they are short, directive recommendation phrases, which results in percentage similarities that appear high when single sentences or fragments are reproduced. Where the 'areas for improvement' sections are very short, a high percentage may be flagged for a very short duplication, such as #10025087 below.

> *Raise pupils' attainment in reading by developing their skills so that they understand new words and phrases, and the deeper meaning of what they have read.*
> - *Ensure that pupils // are developing skills and acquiring knowledge across the full range of subjects.*
>
> #10025087 'area for improvement', October 2017

Here, the first phrase ("Raise…ensure that pupils") is flagged as 62% of the area for improvement being an exact duplication of another area for improvement within the corpus, with the second phrase ("are developing…subjects") a 26% duplication of one other area for improvement. This gives an overall similarity of 88%, despite being only two sentences long. Around a quarter of the areas for improvement with high similarity follow this pattern of being short and almost entirely duplicated elsewhere.

In line with the full reports' findings, the self-referencing approach was seen throughout the 'areas for improvement', and in a pronounced way, particularly for the same lead inspector. #10023088 and #10023065 reports (80% similar) are a prime example of this, where the same lead has used almost identical phrasing, adding additional commentary to make the 'area for improvement' bespoke to the school, but using very similar core text. This is likely an editorial factor, using similar personal linguistic choices that have been proven as acceptable to the quality assurance process. It could also signify that the two schools are very similar in their profile, both good schools in a similar region. A substantial number of reports follow this pattern. It is most pronounced in those with a high similarity report but is a common factor throughout all 'areas for improvement', with many one- or two-word variances only. This occurs for both self-referencing and referencing the language of other inspectors (e.g., #10000915, #10032489 in the following exemplars)

#10036115 has a 88% similarity rating, and self-cites two other reports, #10031497 from June 2017 and #10036133, from earlier that same month by the same inspector. These are a clear example of 'self-referencing' which, when 'areas for improvement' are longer and cover more areas, can subsequently combine to generate a high similarity score. This could be due to the need to cover a recommendation for each of the cited areas of weakness and use of the 'good' descriptors from the Ofsted handbook, or a result of the influence of the quality assurance system. As educational 'jargon' or specialist terminology are restricted to make reports accessible to the public, this could lead to inspectors narrowing their choice of language to expedite the moderation and proofing processes.

The data shows self-referencing is more common in those inspectors with a high inspection output (more than five reports in this corpus) and for HMI (full time employed inspectors), this could indicate inspectors using similar patterns of advice for similar schools (as they have visited many) and using their experience to use the language that has proven beneficial to schools when given in other reports. This could also account for OI duplicating phrases from respected, established HMI. As such, the ‘area for improvement’ were further investigated to see if any phrases or terms held more power or influence than others by using NVivo to spot if trends from frequent authors were appearing in other reports. Threads or replicated language were searched for within the frequent authors, and duplication patterns identified.

Case 1 – an exemplar ‘self-referencing’ from the same lead inspector showing how a lead inspector re-uses long sections of text, even within the most powerful and impactful part of the report. Report #10023088, February 2017 to report #10023065, March 2017. Text is highlighted in bold showing the duplicated portions.

*“I**mprove the quality of teaching and pupils’ learning** to outstanding **by**:*

- ***Continuing the upward trend in pupils’ attainment and progress in reading, writing and mathematics****,*
- ***Ensuring that pupils undertake more extended writing and apply and develop their writing skills in a wide range of subjects****,*
- ***Ensuring greater modification of the curriculum and teaching to fully meet the needs of the most able pupils,***
- *Further reducing rates of persistent absence.”*

Report #10023088, February 2017

“Further ***improve the quality of teaching and pupils’ learning by:***

- ***Continuing the upward trend in pupils’ attainment and progress in reading, writing and mathematics****, particularly concentrating on improving the proportion of pupils who attain above the standards expected for their age,*
- ***Ensuring that pupils undertake more extended writing and apply and develop their writing skills in a range of subjects,***
- ***Ensuring greater modification of the curriculum and teaching to fully meet the needs of the most able pupils****,*

> - *Ensuring that pupils in all classes are helped to see which particular aspects of their work need to be improved, as required by the school's assessment and marking policy."*
>
> Report #10023065, March 2017

Case 2- An exemplar OI referencing earlier HMI 'area for improvement' demonstrating how terms and strings of actions are replicated and become a 'shorthand' way of describing specific improvement actions. Report #10000915, November 2016 to report #10032489, May 2017. Again, with bold text showing the duplicated sections.

> ***Improve the quality of leadership and management by ensuring that:***
> - *Senior* ***leaders regularly make thorough checks on pupils' progress as well as their attainment, and act swiftly to tackle any inconsistencies that exist.***
> - ***Leaders promote the highest expectations for pupils' progress and provide teachers with the ongoing challenge and guidance they require.***
> - *The governing body is provided with the information it requires to regularly and robustly challenge school leaders, holding them stringently to account for the impact of their actions.*
>
> Report #10000915, November 2016

> ***Improve the quality of leadership and management by ensuring that:***
> - ***leaders regularly make thorough checks*** *on the* ***progress*** *of key groups* ***as well as their attainment, and act swiftly to tackle any inconsistencies that exist.***
> - ***Leaders promote the highest expectations for pupils' progress and provide teachers with the ongoing challenge and guidance they require****.*
> - *The governing body strengthens the rigor of its challenge to check the impact of leaders' actions in improving outcomes for all groups of pupils.*
>
> Report #10032489, May 2017

Case 3 – An exemplar high percentage self-referencing across multiple sources showing how a single inspector combines earlier directives from two of their other reports, directly replicating strings of text. Report #10036115, Sept 2017, report #10031497 June 2017 and report #10036133, September 2017. Duplicated text in bold.

> ***Improve the effectiveness of leadership and management by ensuring that:***
> - ***New subject leaders, particularly for English and mathematics, are given the necessary support to help them raise standards and improve outcomes,***

- Action is taken to reduce absence, including persistent absence, especially of disadvantaged pupils,

- Programmes of support for disadvantaged pupils are thoroughly analysed to provide governors with detailed information about their impact,

Improve the effectiveness of teaching, learning and assessment and, as a result, raise standards by ensuring that:

- All teachers share the same high expectations about what pupils can achieve, so that work is appropriately challenging, consistently engaging and well matched to pupils' needs.

- Opportunities are taken to share the best teaching practice across the school.

Improve personal development, behaviour, and welfare by:

- ***Ensuring that systems for promoting positive behaviour in lessons are applied consistently in every classroom.***

Report #10036115, Sept 2017

Improve the effectiveness of leadership and management by ensuring that:

- The curriculum provides far more opportunities for breadth and balance so pupils can consolidate and deepen their understanding, especially in history, geography, science, and religious education.

- ***Action is taken to reduce absence, including persistent absence, especially of disadvantaged pupils,***

- Programmes of support for disadvantaged pupils are thoroughly analysed to provide governors with detailed information about their impact,

Improve the effectiveness of teaching, learning and assessment and, as a result, raise standards by ensuring that:

- All teachers share the same high expectations about what pupils can achieve, so that work is appropriately challenging, consistently engaging and well matched to pupils' needs.

- Opportunities are taken to share the best teaching practice across the school.

- Pupils in every class are given plenty of opportunities to solve problems, deepen their understanding and explain their thinking in all areas of mathematics.

Report #10031497 June 2017

Improve the effectiveness of leadership and management by ensuring that:

- ***Subject leaders, particularly for reading and mathematics, are given the necessary support to help them raise standards and improve outcomes,***

> ***- Action is taken to reduce absence, including persistent absence, especially of disadvantaged pupils,***
> *- **Programmes of support for disadvantaged pupils are thoroughly analysed to provide governors with detailed information about their impact,***
> ***Improve the effectiveness of teaching, learning and assessment and, as a result, raise standards by ensuring that:***
> ***- teachers make better use of assessment information to plan lessons which are appropriately challenging, consistently engaging and well matched to the needs of all pupils.***
> *Improve behaviour in lessons by:*
> *- Reducing low-level disruption and **ensuring that systems for promoting positive behaviour in lessons are applied consistently in every classroom***

Report #10036133, September 2017

Single word frequencies

Aside from these longer duplicated phrases, a search for individual word frequency was undertaken to see if there were any single words or terms that had equitable duplication across the corpus. Word frequency was run on all 'areas for improvement', to see if there were common terms. Initial scoping using auto-code software on NVivo showed some linguistic trends that can be classified thematically:

Attainment: Able (524), Skills (642), Progress (512)
Children: Disadvantaged pupils (368), Pupils (1869), Able pupils (441)
Level: Key Stage (366), Stage (396), Years (499)
Actors: Leaders (664), Teaching (545)

Within the top 200 most frequent terms for the 'area for improvement' sections only, using the same stop words as when searching the full reports (see Ch 3), progress was cited often (512), and linked to signifiers such as 'outstanding progress' (7 times), strong progress (38) rapid progress (114) and good progress (161) as required actions. In contrast, 'improvement' was cited less frequently (210 times) improving pupils (64 citations), improving teaching (38), improving plans (33) improving outcomes (30) and improving communication (10). all relatively infrequent compared to progress.   Rapid improvement (20) and recent improvement (15) indicates more urgent directives to those settings who have recently changed grades, with ‘school improvement planning’ only cited 11 times. This would seem to indicate that the directive for improvement is implied rather than directly stated as an action.

The semantics of the area for improvement are critical, as even a very slight change in wording can change the focus of schools and the evidence required to demonstrate those areas have been addressed. For example, “Improve planning to ensure all pupils are challenged…” requires evidential proof of better planning, whereas “Improve levels of challenge by improving planning…” requires evidence of improved challenge as well as improved planning on subsequent inspection. Due to this, although individual word counts were noted, and patterns identified, the contextual placement of signifiers and qualifiers was also tracked through the frequency data. In practice, this meant two-word phrases were used in most cases to identify not just single terms but the implied directives. For example, this enabled the separation of ‘teaching’ from ‘teaching assistants’, highlighting the adjective rather than the noun.

Concrete actions such as assessment (1121) are cited frequently. These represent actions that can be evidenced by the school through established practices, with written assessment evidence or data that can be recorded. Imprecise directives such as consistent (925), challenge (991), plan (657), or expectations (659) that are more subjective can only be evidenced through leadership judgements, records of monitoring, or actions at the point of re-inspection, which are much more difficult for schools to exemplify.

Actors and Subjects

To interrogate the word frequencies and phrases for the location of power, and identification of who the reports described and gave actions to, texts were searched for actors and subjects of the report and area for improvement. ‘Teachers’ were the most cited actors within area for improvement sections (1128 citations), almost double the next most frequently cited actors ‘Governors’ (683). This implies that teachers are the ones who are expected to act on, complete, and be affected by the requirements. Middle leaders (224) and Senior Leaders (75) feature much less frequently, either in the stem of an area for improvement or implied in the phrasing of the action as a monitoring activity. In the area for improvement below, for example, wording relating to teachers gives a clear indication that teachers are required to complete an action, whereas parts relating to leaders and governors cover tracking of data and impact. In these examples, teachers – the actors are to implement and be accountable for the required actions (bold added for emphasis):

> *Accelerate pupil's progress by making sure that:*
> *- **Teachers** use what they know about pupils' knowledge and understanding to plan learning activities that meet their needs, especially for the most able and least able pupils,*
> *- Pupils understand what they have read.*
> *Improve **leadership and management,** including **governance**, by making sure that:*
> *- **Leaders** effectively track pupils' progress from their starting points in order for **teachers and leaders** to use this information to quickly identify and support pupils who are falling behind their classmates with similar abilities,*
> *- **Middle leaders** use the results of their checks on the effectiveness of their actions to quickly identify and successfully address areas of weakness.*
>
> Report #10019995, November 2016

'Teaching assistants' are cited less frequently (77) and references almost exclusively tied into a teaching action, e.g., "teachers and teaching assistants should…". Within the 'areas for improvement', the object of the directives is almost always pupils. Groups of pupils differentiated in descriptions include 'Early Years' (367 citations), 'Pupil Premium' (185), 'disadvantaged pupils' (365), and 'able pupils' (464). These groupings allow for action to be tailored to a sub-group of pupils which is discrete enough to secure an evidential impact, without being so restricted as to make individual pupils or teachers observable from reports.

Descriptions of the actions to be undertaken for pupils vary widely, and include implementing change (e.g., assess, provide, develop, give, help, enable, track, move, encourage), targets (e.g., improve, raise, deepen, progress), and detailed separation signifiers, often by ability (e.g., lower-attaining pupils, middle ability pupils, more able pupils). As subjects rather than actors, pupils are 'receiving' the required actions, and are anonymised by their nominal determination as a group. Where an action relates to a target, the children who are named during the writing of the 'area for improvement' are rarely the same group of children on re-inspection, which implies an assumption that these nominal groupings retain similar characteristics over time.

When area for improvement sections describe discrete subjects and objects there are limiting factors on a school's ability to demonstrate change over time, as cohorts vary widely, and performance (progress or attainment data) can be skewed by contextual factors. When the periods between inspection are long durations (as much as 4 years) then cohorts are often not directly comparable or reflective of the historical patterns. This could be a reason why ‘progress’ and generic objects and subjects may be more commonly used within the ‘area for improvement’ than specific outcomes for a particular Key Stage or year group who have been identified during inspection.

Curriculum areas

A wide range of specific curriculum areas is cited in the areas for Improvement. Mathematics is the most cited in the corpus (1028 citations), with Writing (982), and Reading (641) not far behind (Figure 6). This is not surprising given these areas are the focus of statutory tests for primary school children. Other curriculum areas are mentioned far fewer times, for example, the next mandatory statutory area, phonics, is only mentioned 146 times, and science only 134 times.

Other subject areas were even fewer still. Spiritual, moral, social and cultural (SMSC) cross-curricular elements were cited; spiritual (8), moral (8), and social (14) were cited as infrequently as arts subjects, whereas cultural (53) was cited at a level closer to history (80) and geography (76). Where these are cited, they appear as areas that have been given in the ‘areas for improvement’ section the inspection framework has very little to say about individual curriculum subjects outside of English and mathematics.

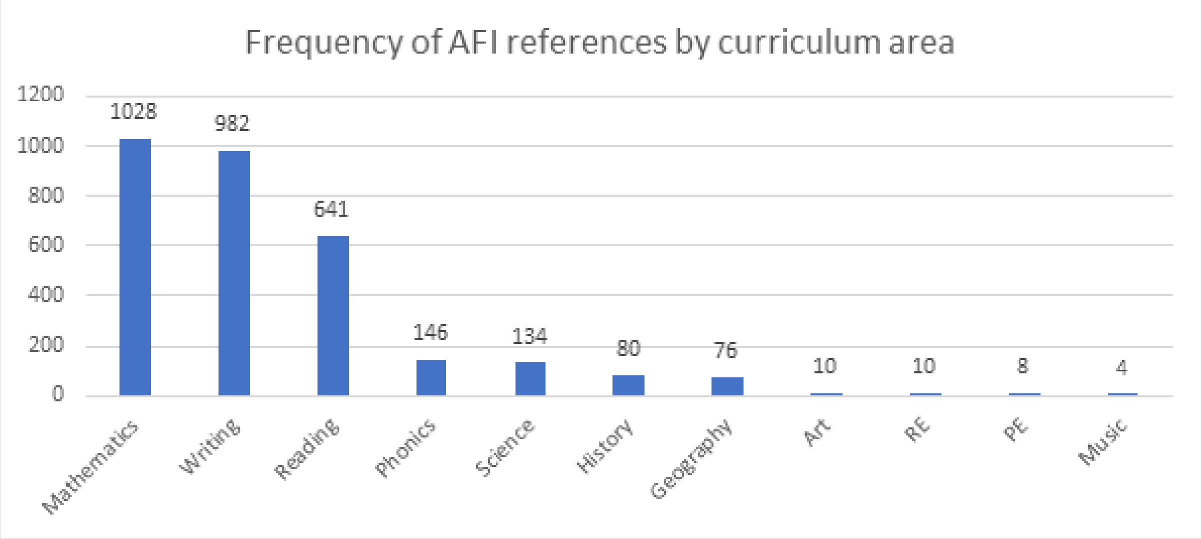


Figure 6 Graph showing the frequency of references to different curriculum areas within ‘Areas for improvement’ sections of the 2017 primary section 5 Ofsted reports.

It is possible to make requirements in relation to curriculum improvement without referring to particular subjects. As an example of this, literacy and numeracy terms were investigated as stand-alone elements (Figure 7).

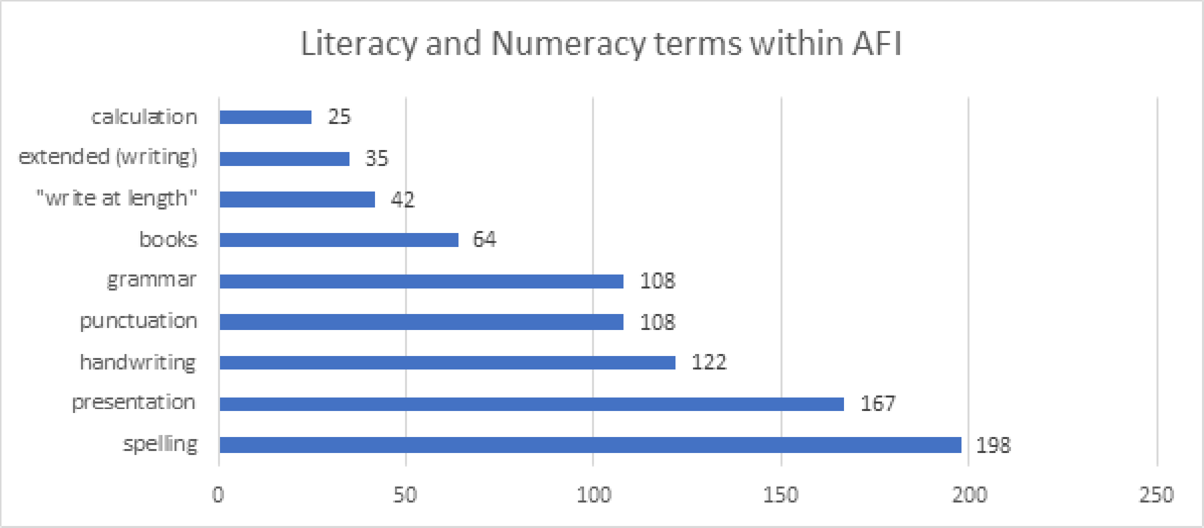


Figure 7 Graph showing the frequency of references to literacy or numeracy elements within 'area for improvement' sections of primary section 5 Ofsted reports in 2017

Measurable skills such as spelling, presentation and grammar were mentioned more frequently than foundation subjects; spelling (198 citations) was more frequent than phonics (146) or science (134). This may be because English spelling, grammar, and punctuation are formally measured and reported upon using national comparative data. Other elements related to literacy, such as the 'ability to write at length' (42 citations) or 'extended writing' (35) are given within 'areas for improvement' elements, however, these do not feature in nationally comparable statistics, and so reflect may reflect a particular pedagogical choice. Equally, writing skills (91 citations) and mathematical skills (99) are frequently given as instructions in the area for improvement, and what is meant specifically is open to interpretation; the measurable components of mathematical skill, such as reasoning (60) or calculation (25) are less frequently cited. This could be due to the time limits of an inspection not allowing enough time to pin down a specific element of the curriculum, amendments to language during the QA process, or the tendency to give more general advice to lower graded schools. It could also be inspector preference for a more general target, or a lack of simple ways to describe a particular curriculum element. Those curriculum areas most frequently cited in this corpus correlate with the focus of SATs (Standard Assessment Test) papers, which could be influencing the choice of language.

As indicated above, Mathematics was the most frequently cited curriculum area. More than half of the area for improvement in each month of 2017 included instructions to improve mathematics (Figure 8).

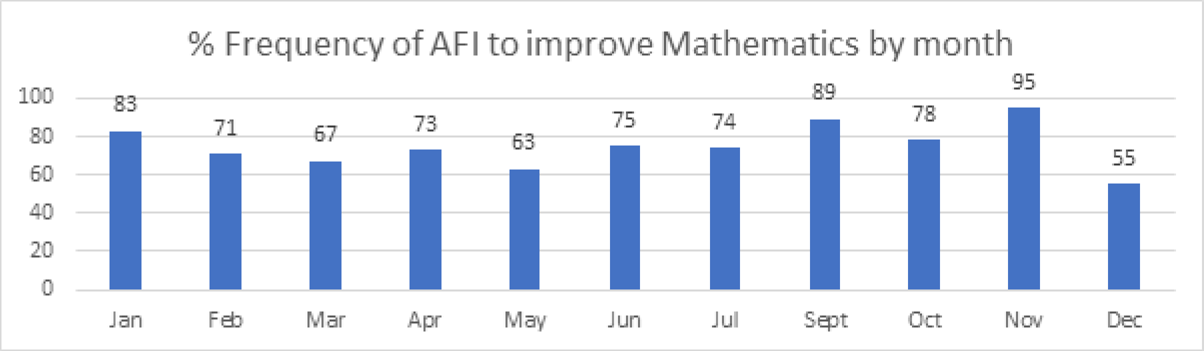


Figure 8. A graph showing the proportion of areas to improve that included a request to improve Mathematics for each month across the year 2017

As Maths is subject to formal, external measurement (SATs), with results that are publicly available, this subject area was selected as a proxy to explore correlations between the language of reports and outcomes for pupils.

4.9 Investigating Mathematics from 'area for improvement' to impact on outcomes

Due to its citation frequency in Areas for improvement, mathematics was selected as a means to track the relationship between language used frequently in 'areas for improvement' and specific outcome data, to investigate whether the language used in Ofsted reports correlates to any subsequent measurable improvement in outcomes (RQ2). To this end, external, validated data, SATS (standard assessment test) outcomes, for the schools within the corpus was used that spanned before and after the inspection and reporting period (i.e., 2016, 2017, 2018, 2019). A new SAT had been launched in 2016 and the government was publishing outcomes annually that showed progress and attainment in reading, writing, mathematics and 'English grammar, punctuation and spelling' (EGPS) by the end of Key Stage 2.

Reports with an area for improvement that included the term 'mathematics' (or lexical variations, using the NVivo automated index) were put into a sub-set. The published mathematics data for the corresponding schools was sourced from the department for education primary school performance tables. The score from the last SATS test before the point of inspection (2016) were used as a baseline and collated for 2017 (the year of the inspection report), 2018 and 2019 (the point of re-inspection for most), using both average scaled scores for attainment and DFE progress measures for mathematics. This covered the formal external published data for mathematics before the inspection, which would have been seen at the point of inspection, through to when the improvement request for mathematics was given, and then scores in mathematics after that point through to the point of re-inspection.

As ‘Mathematics’ was so frequently mentioned in improvement areas, the initial sub-set consisted of 691 reports (half of the full corpus). However, when entering the associated mathematics data for each school, some issues arose. Due to the size of some schools, no data was published. When the number of pupils in a class is so small that publishing data would identify individual pupils, those data are not entered into the government database. For these schools, either ‘x’ indicates for missing or ‘supp’ for ‘suppressed data’ (supplied but would identify individuals because of SEND or similar categorisation) is indicated. Hence, there were gaps in the data set that would make correlations between language and improvement difficult to investigate for those schools. Those who had no clear starting point (2016) or end point (2019) for either progress or attainment due to these missing or suppressed data factors were discounted from the set, which reduced the sample size to 587 (see Appendix 2).

The remaining sub-set was deemed sufficient to investigate the correlation, as it represented roughly 42% of the original corpus. This group, due to the nature of the exclusion parameters, included fewer small schools, but included comparable representative geographical, contextual (urban/rural/coastal), and academy/maintained proportions. Those schools who had academized since the initial inspection could be included, as the government data site automatically linked the old and new Unique School Identifier numbers, and given that the buildings, staff, and pupils remained the same, the mathematics data from the academized school was used, as the data related to children who would have been in the school at the point of the last inspection and who had benefitted from any changes. To analyse the correlation between areas for improvement and impact on mathematics outcomes over time, the two main government measures were used, progress figures and average scaled score (attainment).

Progress

Progress measures are an average for the school and reflect progress between the end of Key Stage 1 (as measured by SAT scores) and the end of Key Stage 2. A score of 0 indicates that children made the progress expected between the end of Key Stage 1 and the end of Key Stage 2. A negative score shows some children did not make expected progress, and a positive score shows some children made accelerated progress. A handbook was published by the Department for Education (DFE) when this system was launched in 2016 to help parents to understand the figures. The use of this average measure enables schools of different sizes to be compared using a single data point, so here the average across the sub-set and the spread of scores can be used to identify trends over time. As the new testing regime was introduced in 2016 and continued throughout the time frame under investigation here, scores are comparable over the four years (see figure 9).

For those schools who had been given an ‘area for improvement’ relating to mathematics in 2017, their average progress in the previous year (2016) was -0.97, below the national average of 0.0 and significant, as tolerance was set to roughly 0.1 for this period. This equates to 393 (67%) of schools in the sub-set having scored below national average for progress in 2016. In 2017, the national average for progress rose to +0.1. During this year, the sub-set were inspected and directed to improve mathematics. By the end of the inspection year, maths progress had improved, but within the sub-set remained below national average rates of progress, at -0.55. The table in Figure 9 shows the progress of the sub-set - those given an area for improvement for maths - relative to national averages over the period 2016 to 2019. For each year tolerance remained roughly 0.1 and confidence intervals remained stable.

	2016	**2017**	**2018**	**2019**
National progress in Mathematics	+0	+0.1	+0.1	+0.1
Those given maths as an area to improve - average progress	-0.97	-0.55	-0.17	-0.25

Figure 9 A table comparing average progress in mathematics for the group given an area to improve for mathematics compared to all schools nationally that same year.

Between 2016 and 2018, 87 schools (15%) in the sub-set had increased their progress in mathematics to above national average. The highest progress score remained stable at an average of +7.9 and the lowest progress in the set improved from -13.5 in 2016 to -9.3 in 2019, which implies that the progress improvements were strongest in the lowest performing settings. The relative stability of the bell curve distribution indicates an overall positive impact, with the change in the average progress measures of this group from 2016 to 2019 being an upward trend, improving by an average of +0.73, which is 7 times faster than the national rates of improvement, and closing the gap to national average score by +0.63. As most schools nationally would also have been inspected between the 2016 and 2019 national data sets being generated, the use of national data as a comparison mitigates the influence of inspection itself triggering improvement rather than any specific terminology within the report itself.

Of those schools that improved progress the most, the majority had improved from very low mathematics scores, and their areas for improvement appear to have few pedagogical directives specific to mathematical learning.

> *Improve teaching further by:*
> *– ensuring that presentation in mathematics books is as consistently neat as in English books.*
>
> Extract from the 'area for improvement', report #10031717

> *Improve outcomes for pupils in reading, writing and mathematics, particularly in key stage 2, by ensuring that:*
> *– teachers have consistently high expectations for what pupils can achieve*
> *– pupils are clear about what they are learning and for what purpose*
> *– there is more challenge for most-able pupils to achieve the higher levels and work at a greater depth.*
>
> Extract from the 'area for improvement', report #10032972

Within the top 100 schools in the sub-set that had most improved progress, only four were above national average when they were given the area for improvement of mathematics in 2016: #10031717 (+0.7), #10033047 (+0.6), #10037754 (+0.5), and #10036473 (+0.2). This indicates that, for most schools, improvements in progress reflected a closing of the gap to national from a low base: 60% of the schools given an area for improvement for maths improved faster than national, with an average uplift of +0.73, which implies that schools that do improve do so rapidly and significantly, and where improvement is not rapid, it is in line with national averages.

The grading of those schools who made improved progress in maths was strong: 54% graded Good or better, only 24% RI, and 22% not previously graded. This indicates a strong correlation between a previously good or better judgement and the ability to make improvements in progress in a specific curriculum area. When linked to the earlier findings that good schools are given areas for improvement that include more actions that require monitoring and more direct links to regulatory language, this implies that for those schools that are already performing well, and who are required to track and robustly measure mathematics, this has a positive impact on outcomes in mathematics.

Of those who did not improve progress in mathematics over the period 2016 to 2019, only 37% were previously Good or better. 35% were previously RI, and 28% were not previously graded. These correspond to the schools given more generic areas for improvement that refer to broad themes rather than specific curriculum elements. This suggests that the more ambiguous an area for improvement, the less likely a school is to have improved progress outcomes in that part of the curriculum.

	2016	2017	2018	2019
National progress in Mathematics	+0	+0.1	+0.1	+0.1
AFI Maths: average progress	-0.97	-0.55	-0.17	-0.25
National average scaled score in Mathematics	103	104	104	105
AFI Maths: average scaled score	101.67	102.98	103.49	104.1

Figure 10 A Table comparing average scores in attainment and progress for the group given an Area for improvement (AFI) to improve mathematics compared to all schools nationally that same year.

Attainment

In 2016, when the new testing measures were announced, a complex ‘scaled score’ method was introduced as a measure of attainment (see Figure 10), which set an ‘expected standard’ for each subject. In 2016, for mathematics, the national average scaled score was 103. The sub-set of schools with an improvement area for mathematics had an average scaled score of 101.67, with 362 (62%) scoring below national average according to SAT data for that year. As a measure of attainment, the score represents the proportion of pupils that reach ‘expected standard’ in KS2 SAT tests of mathematics, and it takes into consideration SEND and cohort variance across schools. Tolerance was set to 5% (or +/- 6 scaled scores) in 2016, due to it being the first year that this type of data was produced. By the end of 2017, during the inspection cycle which generated the improvement area of mathematics for this sub-set, only 319 (54%) of schools were below national average. Tolerance was drastically reduced by the DFE over the period under investigation here as results remained stable, so a gap to national of +/- 1 is more significant in 2019.

Over this period, those with mathematics as an ‘area for improvement’ increased their mathematics attainment by +2.42, which, given the national increase of +2, closes the gap to national over the period by 0.42 yet remains almost a whole point below (Figure 10). This improvement is only marginally faster than national. The lowest scaled score possible is 80, the maximum 120, and for this group, the lowest scores improved from 90 in 2016 to 95 by 2019, and the maximum scores improved from 109 to 115 over the same period.

This indicates that, although averages remain below national, some systematic and sustained improvement in mathematics outcomes had been secured. For those who improved attainment in mathematics, the improvement was from a very low base (2016 average progress of -2, average attainment 100) and the change was significant, a gain of +4.95 by 2019, 2 and a half times more than national improvements in attainment.

When correlated against other variables (inspector status, gender, inspection type, or team structure) to identify trends (RQ3) no correlation was found with either outcomes or being given an ‘area for improvement’ relating to mathematics. Slightly more of those settings that improved were led by HMI (+6%) and by a female inspector (+7%), although this gender balance correlates to the proportions of HMI and female inspectors within the corpus as a whole. T-tests showed no significant correlations for gender, designation, or team size.

When looking at RQ1) How unique are reports? and RQ2) Does the language used in Ofsted reports lead to subsequent improvements in outcomes? in combination with RQ5) Within the area for improvement, are there any linguistic terms that are particularly effective or impactful? The ‘area for improvements’ were investigated for similarity (see figure 11). The ‘area for improvement’ were more similar in those settings who did improve progress, 27.8% compared to 21.7% similar in those who did not improve. This is significant, as the margin for significance within the software is set to 22%. Attainment followed a similar pattern (28.2% who did improve attainment, 22.8% in those who did not).

This finding suggests that there could be common, effective terms that are more successful to improve mathematics. To check this, T-tests were run on the similarity groups separated into ‘did or did not improve progress’ using the national scores for each school between 2016 and 2019. The T-test showed there was a significance ($p<0.01$) in similarity. The mean similarity for those who did improve progress was 27.82%, compared to the mean for those who did not (17.66%) showing those who improved progress had significantly more similar areas for improvement.

For attainment, the T tests were repeated, based on the average SATS scores over the same period separating into groups of ‘did and did not improve attainment’. The T tests showed that there was a significant difference ($p<0.01$) in similarity. The mean similarity for those who did improve attainment (28.7%) was significantly higher than the mean similarity for those who did not (22.4%). This shows that the area for improvement were statistically more similar for schools who improved attainment (See Appendix 14-16 for full tables of results).

These two together suggest that there is terminology that when given as directives to improve, helps schools to improve not only the progress pupils make, but their overall attainment in mathematics. This shows that improvements do not simply show an improvement in test results for single cohorts of pupils (attainment0 but reflect improvement in mathematics for pupils over time and from different starting points (progress). This also shows that a more bespoke or unique description of what to improve does not correlate to improved outcomes for pupils in either mathematics progress or attainment.

Figure 11

A table comparing the improvement profiles and similarity averages for those given an Area for Improvement (AFI) for mathematics, the 2017 corpus and all schools nationally.

Average	**Whole Corpus/ National** (1391)	**All Maths AFI** (586)	**Those who improved progress** (n=353)	**Those who did not improve progress** (n=248)	**Those who improved attainment** (n=277)	**Those who did improve attainment** (n=309)
Progress	+0.1	+0.73	+2.65	-2.04	+2.73	-1.07
Attainment	+2	+2.49	+3.88	+0.25	+4.95	+0.16
Similarity of Area for Improvement (%)	21.1	25.4	27.8	17.66	28.2	22.8
Similarity of full report (%)	55.5	55.9	55.9	55.8	56.2	55.6

‘Area for improvement’ language leading to most improvement

The apparent correlation between terms in similar areas to improve and levels of improvement (RQ5) was investigated further to identify whether there are any linguistic terms where the correlation is particularly strong. Two schools with high levels of improvement in maths are used below for illustration.

School	2016		2019	
Graded: RI #10019298	Attainment	Progress	Attainment	Progress
	90	-13.5	105	+15.4
Graded: Good #10031364	Attainment	Progress	Attainment	Progress
	104	-0.7	115	+7.4

Figure 12, comparison of two outlier schools

The improvement shown for #10019298 was from a very low point in 2016 when it was graded 'requires improvement'. In fact, for the 15 most improved schools, all but two have improved from significantly below national averages, with an average starting point of -5 for progress (national = 0) and an average scaled score of 97 (national = 103).

Notably #10031364 school, graded 'good', whose 2016 progress score was 1 point above national (103). This is what was referred to at the time as a 'coasting' school, where children attained well but did not make the full progress of which they were deemed capable. By 2019, its progress and attainment far exceeded the national averages (105, progress +0.1). In both these cases, the areas for improvement include discrete pedagogical instructions, requirements to complete actions that would directly influence curriculum choices, e.g. citing reasoning and problem-solving specifically.

Report #10019298 reads:
Improve the quality of teaching, especially in reading, writing and mathematics, by:

> *- developing and deepening pupils' mathematical understanding through activities which allow them to apply their **skills, problem-solve** in a range of contexts, **reason** and justify their answers*

Similarly, Report #10031364 reads:
Improve outcomes so that more pupils reach the highest standards expected for their different ages by making sure that:

*- pupils are given frequent opportunities to use **reasoning** and **problem solving** in order to deepen their understanding, especially in mathematics.*

On further investigation using NVivo, the areas for improvement of those schools that did improve mathematics cited problem solving 67 times. In a couple of notable cases -

#10025640 and #10042651 - problem-solving and reasoning were requested twice in the same area for improvement section:

> *Improve leadership and management by:*
> *– improving pupils' achievement in mathematics through opportunities to develop their* ***problem-solving and reasoning*** *skills*
> *Improve the progress of children in the early years by:*
> *– making sure that children have opportunities to use their* ***reasoning and problem-solving skills*** *in mathematics.*
>
> Report #10025640

> *Strengthen leadership and management by:*
> *– closely scrutinising pupils'* ***reasoning and problem-solving skills*** *in mathematics, to ensure consistent challenge for the most able pupils.*
> *Strengthen teaching, learning and assessment by teachers:*
> *– consistently ensuring that pupils practise their calculation skills through* ***reasoning and problem-solving*** *in mathematics*
>
> Report #10042651

These terms were relatively new to Ofsted reports, as they refer to particular sections of the reformed SATs tests that were introduced in 2016. It would be reasonable to assume therefore that they were in common use by teachers during this time. The Ofsted Handbook update for this period refers to reasoning, problem and calculation once each (#175).

> *In the mathematics lessons observed, through discussions with pupils and scrutiny of their work and by reviewing curriculum plans, how well teaching:*
> *- fosters mathematical understanding of new concepts and methods, including teachers' explanations and the way they require pupils to think and reason mathematically for themselves,*
> *- ensures that pupils acquire mathematical knowledge appropriate to their age and starting points and enables them to recall it rapidly and apply it fluently and accurately, including when calculating efficiently and in applying arithmetic algorithms…*
> *- …enables pupils to solve a variety of mathematical problems, applying the mathematical knowledge and skills they have been taught.*
>
> "Inspecting the impact of the teaching of Mathematics" Ofsted, 2017, #175

Across all area for improvement sections in the corpus, problems (or its lexical equivalents using Turnitin) appear in 704 times (51% of all areas for improvement). Such a common term, appearing at this frequency in that year, and which has specific pedagogical meaning in Mathematics prevalent at the time, could directly influence curriculum choices in many schools.

To investigate the link between inclusion or discounting of core mathematical terms and the associated progress in mathematics schools make, a single factor Anova was run. This isolated groups using the variable of progress in mathematics from 2016-2019 in schools where the area for improvement part of the report

- includes the single word ‘x’
- includes the combination of words x **and** y
- **does not** include the word ‘x’

The number of groups is initially significant, as there were 8 key words, linked to the SATS paper from 2017 which were showing as frequent words and which seemed to correspond to higher average progress for reports including those words compared to national progress: Problem (+1.04), Reasoning (+1.05), Accuracy (+2.02), Number (+1.17), Calculation (+0.37), Most Able (+0.76), Challenge (+0.83) and Expectations (+1.18). National average progress for the same period was +0.1.

To see whether these single words were statistically significant, or if they were significant in combination or compared to areas for improvement without those words, a comprehensive Anova was generated (See Appendix 15) which showed p=0.035 for the full group, indicating significance across the groups. T-Tests on each variable were run, and the full list of 107 pairs of groups that showed as ‘significant’ is shown in Appendix 16. For the purposes of this analysis, a selection has been chosen to exemplify the findings.

The progress data for schools whose reports had an area for improvement including the term ‘mathematics’ had a mean progress score for mathematics more than 7 times higher than the national progress score for the same period. Within that group, those areas for improvement that included the word ‘problem’ had progress that was more than 10 times higher than national. The term ‘reasoning’ corresponds to progress scores also 10 times higher than national. Within the corpus of those with a mathematic area for improvement, those without the words problem or reasoning have a mean progress score of five times above national. This was further investigated as part of RQ5 whether there are some terms that are more effective and impactful. This data suggested there were terms that when used, corresponded with better average progress outcomes in mathematics.

In this case, where a school is told to improve mathematics, then the outcomes in mathematics are statistically likely to improve faster than national, and when specific key words related to the SATS tests upon which outcome measures are based are used within the area for improvement, then improvement is likely to be even more significant, and those SATS specific words together appear to have a cumulative improvement effect. Attainment seems to follow a similar pattern of increased impact (see figure 12) however, for this investigation, it was necessary to focus on a single measure of impact, and progress was chosen.

There are generic education terms, not specifically mathematical, that are frequently replicated across these areas for improvement, including ‘expectation’ ‘most able’ and ‘challenge’. These were included as comparisons to the specific mathematical words. A text search was done for the other common descriptors and their lexical equivalents from the 2016 SATS paper question areas; number (46), calculation (22), and accuracy (30) (see figures 12, 13 and 14, appendix 15 for Anova and 16 for T-Tests).

Figure 13 A table comparing Mathematics SATS outcomes from 2016-2019 for those schools given areas for improvement (AFI) including the words ‘problem’ and ‘reasoning’ in 2017

	National (21,253 schools)	**Those with AFI Maths** (n=586)	**Including *problem*** (n=143)	**Including *reasoning*** (n=133)	**Neither *problem* nor *reasoning*** (n=402)
Mean Progress	+0.1	+0.73	+1.04	+1.05	+0.55
Mean Attainment	+2	+2.49	+2.34	+2.56	+2.4

Figure 14 A table comparing average Mathematics SATS outcomes from 2016-2019 for those schools given areas for improvement including the words 'number' 'accuracy' and 'calculation' in 2017, and their associated similarity reports

Average	**Whole corpus / national** (1391)	**All Maths AFI** (n=586)	**Including 'number'** (n=47)	**Including 'accuracy'** (n=30)	**Including 'calculation'** (n=22)
Progress	+0.1	+0.73	+1.17	+2.02	+0.37
Attainment	+2	+2.49	+2.89	+3.3	+1.68
AFI similarity (%)	21.1	25.39	26.6	32.6	27.8
Report similarity (%)	55.5	55.89	56.3	55.7	54.4

Figure 15 A table comparing SATS Mathematics outcomes 2016-2019 for those schools given areas for improvement (AFI) including the non-mathematics specific words 'most able' 'challenge' and 'expectations' in 2017, and their associated similarity reports

	Whole corpus / national (1391)	**Maths AFI** (n=586)	**Including 'most able'** (n=377)	**Including 'challenge'** (n=194)	**Including 'expectations'** (n=206)
Progress	0.1	0.73	0.76	0.83	1.18
Attainment	2	2.49	2.5	2.4	2.75
AFI similarity (%)	21.1	25.39	28.2	29.5	31.7
Report similarity (%)	55.5	55.89	56.5	56.9	56

The T Tests following the Anova showed 107 areas of statistical significance ($p<0.05$). These can be seen in full in the appendix. Some of note are the mean progress for the areas for improvement that included the word accuracy (+2.02) compared to those without the word accuracy (+0.66) ($p=0.018$) and those including the word expectations (+1.18) compared to without the word expectations (+0.48) ($p=0.009$). For these two words in isolation, there is a positive association with improved mathematics progress.

Between the groups, the means for accuracy (+2.02) compared to calculation (+0.37) proved significant ($p=0.004$) and also accuracy (+2.02) compared to most able (+0.76, $p=0.003$). Within groups, all but 4 out of the 28 possible combinations of two of these words within the same area for improvement have means that are above the national (+0.1) and some significantly so. 'Accuracy and challenge' as a group has a mean progress of +2.86, 28 times higher than national, 'number and reasoning' has a mean progress of +2.47, 24 times higher than national. Figure 16 with all of the permutations shows that 86% of combinations of key words results in a mean progress that is higher than national.

Where words are combinations of mathematical terms only (accuracy, calculation, number, problem, reasoning) the mean is +1.05, ten times above national. Where there is a combination of mathematical and non-mathematical (challenge, expectations, most able) the mean is +1.27, 12 times higher than national. However, where there are only combinations of non-mathematical terms, the mean progress is +0.95.

Figure 16 A table collating the T-Test P value results comparing means of Mathematics SATS progress between 2016 and 2019 showing where combinations of words are showing significance.

	Accuracy *(n=30)*	Calculation *(n=22)*	Challenge (n=194)	Expectations (n=206)	Most Able (n=377)	Number (n=47)	Problem (n=143)	Reasoning (n=133)
Accuracy *(n=30)*		3	2.86	2.5	2.12	**-0.6**	1.43	2.33
Calculation *(n=22)*	3		**-0.15**	0.98	**-0.19**	0.36	0.29	**-1.1**
Challenge (n=194)	2.86	**-0.15**		0.86	0.93	0.55	0.84	1.22
Expectation (n=206)	2.5	0.98	0.86		1.06	1.07	1.75	1.83
Most able (n=377)	2.12	**-0.19**	0.93	1.06		1.09	1.13	1.16
Number (n=47)	**-0.6**	0.36	0.55	1.07	1.09		1.17	2.47
Problem (n=143)	1.43	0.29	0.84	1.75	1.13	1.17		0.94
Reasoning (n=133)	2.33	**-1.1**	1.22	1.83	1.16	2.47	0.94	

‘Area for improvement’ language leading to the least improvement

For those schools who did not improve, some were notable outliers. For example, #10042802, a school graded as Requires Improvement, fell from +0.4 progress in 2016 to -9.3 in 2019, with an attainment score of 104 in 2016 dropping to 96 in 2019 (-9.7 progress, -8 attainment). In this case, the area for improvement were specifically focused on leadership and are generic; they do not relate to any specific pedagogy or area of statutory testing. Equally, report #10036618, another RI school, fell 11.1 points in progress, with a -7 average scaled score, and #10024104 declined -9.3 points for progress with a -6 average scaled score.

An example of their area for improvement is given below:

> *Raise achievement across the school by ensuring that:*
> *– there is a focus on rapidly raising expectations and improving outcomes for pupils, particularly in writing and in mathematics*

Report #10042802

> *Improve the quality of teaching and learning across the school by ensuring that:*
> - *work is more carefully matched to pupils’ abilities in writing and mathematics, particularly the most able and most able disadvantaged pupils, so that they are challenged to achieve the highest standards*
> - *pupils are provided with activities to promote their problem-solving and reasoning skills in their mathematics work.*

Report #10036618

> *Improve pupils’ outcomes by:*
> *- providing more challenge for the most able pupils so they are able to attain a greater depth in their learning.*
> *- Helping pupils in key stage 1 make swifter progress to reach national expectations in reading, writing and mathematics.*
> *Further develop the curriculum, especially in science, so that pupils gain subject specific skills and knowledge in subjects other than English and mathematics.*

Report #10024104

For these three schools that showed the highest level of decline in mathematics, it is notable that all three have a specific requirement in their area for improvement to improve challenge for the most able pupils in mathematics. Challenge (and its lexical variants) within the areas for improvement was searched and was more commonly cited in those with an area to improve for mathematics (64%) compared to all areas for improvement (57%). More able (and its lexical variants) and expectations were also more present in area for improvement for maths compared to all areas for improvement.

A check of the most common terms found in reports for schools that showed the most and the least improvement showed a positive correlation to outcomes: some words correlated to greater progress, others to improved attainment. No words or phrases correlated to a decrease in progress or attainment. Therefore, it can be said that within this corpus, the use of any mathematical terms from the SATs papers within an area for improvement correlates with an improvement in overall outcomes for pupils, despite this small number of outliers.

Chapter Five

Implications – what do these findings mean for local practice?

The similarity of reports implies that the bulk of the descriptions within a report are common to multiple schools. This suggests that the broad boundaries of what constitutes a ‘good’ school for example, are generic traits able to be described in similar terms applicable to multiple contexts. This matches the inspectorate handbook ideology (Ofsted Handbook, 2017), and most leaders within my practice would accept that finding. The high similarity also suggests that reports are generally consistent across the country and across inspectors. Professional practice shows that some leaders will struggle to accept that despite their personal evaluations of particularly harsh or generous individual inspectors, the bulk of final published report content, and by association, the accompanying judgements, are generally consistent.

That the most unique section of a report is the ‘area for improvement’ – the direct requirements given to leaders to fulfil to reach those generic, common performance outcomes, implies that inspectors have identified the unique context of an individual setting, and are tailoring their advice and requests to a single specific school. This means that both leaders (who feel reports are generic) and the inspectors (who are writing reports specific to a school) are in fact both correct. This knowledge may go some way to alleviating concerns from schools and build confidence in the process.

That there is a direct correlation to the requirement to improve a specific area (mathematics) and subsequent improvement in pupil outcomes will be important news to both school leaders and the inspectorate itself – as there is no current evidence base that tracks associated impact in this way. School leaders who already put high regard onto the area for improvement section because of re-inspection protocols, will now be able to scrutinise the exact wording and may change their discussions with inspectors to lobby for more specific language in the areas for improvement in future inspections.

Discussion of findings – qualitative analysis and evaluations
Research question 1 (RQ1): How unique are Ofsted reports?

The analysis of 1391 full Ofsted reports on primary schools published in 2017 contained on average 50% duplicated content. In approximately 10% of reports, 25% of the content was directly duplicated from one other single report published that same year, equivalent to a full page of the report. This finding will be notable to school leaders whose perception that reports were overly similar in part informed this study. The finding that such a high level of similarity exists within this corpus raises questions about the extent to which this is common across all years of inspection, and of the purpose of Ofsted reports, and the extent of the influence of reports on school improvement decisions, given the potential correlation to improvement in outcomes for pupils.

Given the restrictions of the reporting framework, requiring inspectors to comment on a prescribed list of specific topics (leadership, pupil premium, outcomes) and the need to evaluate against the established Ofsted handbook criteria, the high level of duplicated phrases and 'inspection shorthand' is a logical side-effect. Feedback from heads and governors that reports had become utilitarian (see chapter 2) is supported by this data. At the time, the final written report would be sent to schools with a feedback questionnaire that contained the question '*does the report content match the feedback?*' This enables Ofsted to track the feelings of school leaders regarding the language of reports following inspection. The 'big data' approach taken here (Dastjerdi, 2016) has enabled evaluation of whether the perception of school leaders is warranted. Beyond overall high levels of similarity, trends in duplication were found. For example, in the most similar report, two thirds of the content were duplicated from one other single source, the equivalent of three out of the five pages of text. For this school, only 12% of the report (460 of 4000 words) were bespoke to this school. The quote below shows how similar the phrasing of bullets and sections are. The bold text indicates the section that directly replicates one other source:

> *"Teachers generally* ***match work accurately to pupils' abilities, but occasionally work can lack challenge, especially for the most able and children in Early Years.***
> ***- Leaders' monitoring of the quality of provision in English and Mathematics is rigorous and detailed. Subject leaders' skills across the wider curriculum still require further development.***
> ***-Leaders' effective use of the pupil premium funding has resulted in the difference between the progress and attainment of disadvantaged pupils and***

> ***their peers diminishing rapidly****, especially in Key Stage 1. A small difference remains in key stage 2, and only a small number of* ***disadvantaged pupils are working at a greater depth of understanding.***
>
> ***-The leadership of the well-planned curriculum ensures that pupils' visits from across the region fire their enthusiasm for learning. The schools tracking of pupils' progress in*** *creative* ***subjects*** *and in subjects* ***other than English and mathematics is new and*** *not yet embedded.*
>
> ***-The behaviour of pupils is good. They have respect and care for each other and are polite and friendly. They display good manners. There are occasional lapses of behaviour in lessons*** *when pupils lose concentration as the pace of learning slows."*

The front page of report #10037735 showing duplicated text in bold.
(88% similarity report)

This extract is from the front page of the report, where the main findings for each graded section are summarised, and so are additionally influenced by the need for brevity. Even so, the leadership of this school would have had concerns if they had known that so much of what was written about their school was directly replicated from another report by this same inspector, a matter of weeks before. If school leaders have an expectation that a report reflects the specific character, strengths, and weaknesses of their school, this level of duplication could undermine confidence. It implies a view among the inspectorate that schools, their shortcomings, and the actions required to address them are sufficiently similar to be judged 'good' that replication of report content is acceptable.

Beyond establishing the overall level of duplication, the nature of and trends in similarity were investigated. Three main types of duplication were identified: block duplication; template writing; sporadic duplication.

Block duplication

A substantial number of reports followed a block duplication structure, where large sections of text replicated with only minor alterations. These accounted for a large proportion of the high similarity percentage reports (440 reports that were over 60% similar) where a single duplicated block often represented as much as a full page of the total report. This level and style of duplication implies that significant proportions of a report are not unique, in a format and to a degree that would be noticeable to the reader.

These blocks were sometimes due to inspector self-referencing (citing another report the lead inspector had previously written themselves) but also inspector duplication (i.e., duplicating content from a report written by another inspector). These blocks appeared throughout the report, including the front-page and areas for improvement sections, where text is predominantly in bullet points. The front page summarises the main evaluations that have led to each grading, and the area for improvement sections describe the actions the school must take.

The data showed that although there were some inspectors who had higher similarity profiles than others, in general these differences were marginal, and all above the threshold of notable similarity. No gender, designation, team size or authorial trends or variables were observed that led significantly different reports than those reflecting the ‘voice of the inspectorate’. It is more accurate to talk about reports written ‘by the inspectorate’ than any sub-section of it, as this corpus showed that the quality assurance process had effectively removed almost all independent authorial patterns. Only patterns of ‘inspector self-referencing’, where single inspectors replicated their own writing choices over multiple reports, reusing phrases that had already passed through and been accepted by the quality assurance and publishing process were found, and these patterns were subsumed into the voice of the inspectorate, by these choices being replicated by other inspectors. Kesby (2005) describes this reduction of individuality as the ‘performance’ of inspection, whereby the group mimics the power structure, to wear the mantle of authority and assume the ‘voice’ of the inspectorate. This runs parallel to more operational limiting factors such as time, workload, and efficiencies of communication, such as those described in Jaworski and Coupland (1999).

The analysis showed that, as reported by Kogan (1971), HMI have more freedom as illustrated by their reports being marginally more unique. Patterns for the area for improvement were not statistically different for either OI or HMI designation, however. Therefore, if the improvement actions taken by schools in response to the requirements of the report did have an impact on outcomes, then this is attributable to the process of reporting, and by association the work of Ofsted as an inspectorate, rather than any specific inspector.

The skew towards full reports authored by HMI being more unique appears to be due to the use of HMI for potentially contentious inspections or inspections of schools previously graded inadequate. The data showed the largest variations in similarity were for reports of schools graded inadequate schools, 90% of which were led by HMI in this corpus. Reports on these schools are permitted a much larger word count, which may contribute to their greater uniqueness.

Self-citation

The most common similarity pattern was for a report to be made up of multiple single sources each with similarities of around 20%, and these compounded to create a higher overall similarity percentage. Many inspectors adopt similar linguistic patterns, evolving into a self-citation pattern described earlier, where almost identical phrases are used for multiple schools. This pattern exists in other policy analysis and reporting mechanisms and is generally correlated with institutionalised terminology and phraseology (Van Dijk, 2009). It implies that the audience and authors have a shared understanding and shorthand that more efficiently describes actions and evaluations. This suggests that Ofsted reports are not more or less unique than similar reporting mechanisms across other sectors and although the high percentages may appear shocking to school leaders, these are common in this kind of document. This pattern of language embeds government linguistic ideologies into common professional language (see Ch 1.2) which in turn filters into mainstream communication mechanisms. This proliferation of Ofsted handbook terms, which reflect the language of national policy, is the kind of indirect consequence referred to by Ehrens (2014) and correlates to findings in medical fields (Iacobucci, 2018). In broad terms, that language which is embedded by the process of observation and evaluation, and which speaks clearly to the wider public (rather than insider jargon) holds more power, and therefore bigger risk to those institutions that fail to live up to the perceived standard. The more generic and based in accepted policy the descriptor, the more likely the institution is to agree with the judgement, which research has proven is critical to further improvement (see Ehrens, 2015 and Creemers et al, 2007, Chapter 2).

There are phrases that are not required elements, but feature frequently, for example "*The school’s work to promote pupils’ personal development and welfare is good*" (866 replications across the corpus of this exact phrase). This shorthand, compared to a required phrase “*The arrangements for safeguarding are effective*” (1366 replications, including some reports that contain this exact phrase more than once) shows how disproportionately weighted the reporting terminology is on mandatory elements, and how often shorthand phrases are used to efficiently capture evaluations.

Replication blocks of around 10% similarity to another report appear to be the core mechanism for inspector self-referencing, where inspectors use the same phrasing or way of describing a particular aspect across multiple reports. Initially, the assumption had been that the ‘voice’ of the author came from the lead inspector, but the analysis showed that the ‘voice of the inspectorate’ overrides this. This reflects my own experience, attending inspection meetings as both a school leader and inspector, where the summaries of evidence given during the oral feedback to school leaders uses these ‘shorthand’ terms, which quickly become established lexicon within schools

This form of replication is stylistic, and a response to the highly regimented language required, and the need to issue reports that can be easily read by parents and the wider school community. Avoiding acronyms, educational jargon, and pedagogical language so that the report is easily readable by its audience restricts vocabulary choice, yet some content remains mandatory. For this reason, inspectors appear to have developed a ‘tried and tested’ way of expressing evaluations and use these repeatedly across multiple reports. These are then adopted by other inspectors. This process is described in Chapter 2 with reference to the European context, where the language of inspection becomes a way of both upholding and embedding national policy and influencing the actions and outcomes of schools. Terms that begin as an operational efficiency, and which then influence choices of activity, areas for scrutiny, or which gain status as part of evidence bases, have an important influence within schools. For example, the term ‘outstanding’ now has a specific connotation beyond its literary descriptors, and carries the history of inspection activity, pressure and recognition if used in any description. Similarly, any phrases relating to improvement activity or description of regulatory failure, through very short sentences, refer to a much larger knowledge base and shared understanding by insiders. For example, the phrase “Safeguarding is ineffective” has wide and far-reaching implications and connotations for systems, processes, leadership, and public relations and can significantly change a school’s current and future actions. It is a shorthand recognised by those within the system and increasingly by the wider school community.

Template Writing

A large proportion of duplication was found on the front page of the reports. Here there are considerable limitations on word count, a requirement to write in bullet points, and a limit to a single page of text overall. This appears to lead to less variation of content and the adoption of condensed phraseology. Each report is required to include a critical evaluation summary for each of the graded areas on the front page. If ‘good’ is awarded, there are limited descriptions that match the framework, that can be captured in a bullet point, and that fit within the required word count. Equally, any areas deemed to ‘require improvement’ must be summarised on the front page. This leads to the proliferation of ‘shorthand’ terminology, where set phrases have become indicators of longer descriptions, which is referred to here as ‘template writing’. This implies that the structure of the report itself, which requires particular elements in limited word counts and specific formats, has influenced not only the language of reporting, but also, through this, the actions required following inspection. In turn this will influence the actions of school leaders as they develop school improvement plans based on these aspects of the report.

The commercialisation of the report, making it scalable, consistent, able to be read by a wide audience, and so on, has had affected the levels of power and influence of different terms and required actions. For example, the need to use common terminology within the report has led to a proliferation of terminology that could be described as supporting a style of pedagogy (problem-solving, calculation), something that historically Ofsted tried to limit, so as to not introduce a bias towards any particular teaching approach. When the testing and reporting process has common elements that are mandatory for all children, and with terms becoming mainstream, used frequently in communication to parents, it is easier to pinpoint tested areas that are weaker, and request that schools improve children's performance in those areas using specific terminology.

Whether these schools can better identify and target an action plan or are being directed to 'teach to the test' to show improvement are both reasonable interpretations of these findings, but if the outcome is an improved result on an assessment that the majority of schools accept as a reasonable measure of mathematical ability, then it is cannot be said to be a negative influence on outcomes. Faubert (2009) describes this pattern as a common feature across OECD countries, and an accepted 'norm' within education.

Template writing often uses formulaic 'stems' or 'hooks' that conform to the required content were sandwiched between more unique terms and phrases. This form of duplication accounts for most of the smaller percentage replications (1% to 7%). For example, the extract below shows bespoke sections in between the duplication (in bold):

> ***The leadership team monitors teaching and learning well.*** *This has helped leaders to evaluate the effectiveness of the curriculum.* ***The curriculum is well designed as it enables pupils to*** *experience a range of topics, including spending appropriate amounts of time exploring each other's opinions. Such work* ***enables pupils to develop their spiritual, moral, social, and cultural understanding well.***
>
> Report #10025177

This differs from self-referencing as the sections often jump from one topic to another, with shorter duplications followed by a unique ending, or a bespoke opening closed with a formulaic evaluation. This pattern was found in all reports within the corpus to differing extents and contributed to around a quarter of all duplication.

This appears to be an in-built consequence of writing to the report requirements, using the given structural limitations and language limitations, leading to the recognisable ‘Ofsted voice’. Most template writing includes direct replications of terms from the Ofsted handbook.

Former chief inspectors have tried to recapture the historical role of the HMI (see chapter 2.3) as a more humanist interpreter of national policy, but the increasingly bureaucratic and high stakes inspection process, particularly since the introduction of academies in 2002, has contributed to an increase in government terminology being found within inspection processes, as they are increasingly embedded within the terminology of the Ofsted handbook itself. Although the activities undertaken by inspectors on site may be more bespoke, the corpus of written reports shows that the narrative produced is highly standardised, with terminology limited to those elements universally accepted as central to evaluation and that have a widely accepted meaning among the target audience.

When describing a school’s performance against the Ofsted framework, it is difficult to use original expressions without breaking restrictions, risking ambiguity, or exceeding word limits. These limitations may also affect the choices of activities inspectors complete when on site. For example, viewing a heart-warming assembly would better enable an inspector to capture the uniqueness of a school, but the inclusion of this within the tight timetable of an inspection visit is often over-ridden by the need to observe mathematics, hear reading and so on, which all have required reporting conditions. Therefore, an efficient inspection has to prioritise those elements that are mandated by the reporting framework, which impacts not only which elements are included within an inspection, and are subsequently reported on, but also therefore which elements can be identified as areas for improvement. This will shape the focus of the school’s improvement activity and in turn the content of the next inspection.

The ‘utilitarian performativity’ of inspection (Ozga and Lawn, 2014), where the need to become expert at being evaluated in this manner, due to the public and intended consequences of inspection, have been shown in the wider literature to steer school leader’s choices towards national policy. The requirement for standardisation, at the heart of Ofsted’s creation (see Ch 1.2) has been the instigator of this reduction in language and move towards measurable performance indicators that conform to a business model of quality assurance. These findings reflect the commodification of inspection, a service of equitable quality and commercial standardisation with predictable outcomes (see Ch 2.3). The reduction in terms and descriptions implies a similarity between settings and across the wider spectrum that may not accurately capture the uniqueness of individual schools. Additional influences, such as changing the internal language within schools, leading to increasingly standardised practices, and to some extent 'teaching to the test' are those elements which feature frequently in school leader conversations.

In conversations across my professional practice, school leaders frequently comment on how risk-averse and inspection-compliant they are required to be in order to maintain positive public evaluations. In summary, the instrumental nature of reporting has led to a reduction in the terms used to both describe and drive school performance. These distilled terms then impact on leaders' choices, which then impact on future inspection activity. This cycle of reductionist activity has led to the current situation, where inspection reports are a collation of common indicative phrases relating to a small number of common foci. It is upon these listed areas that school performance is judged, by selecting from an accepted bank of descriptors, rather than describing individual circumstances and bespoke provision of education.

Heavy duplication in all these areas implies that Ofsted thinks that schools have not only very similar profiles but similar tasks to complete in order to improve their quality. This is an important finding. Working this through to a logical next step, it would theoretically be possible, using this software and a larger sample of only the 'area for improvement' sections of reports, to identify actions requested by Ofsted, and rank them in frequency order. This would give a blueprint, which could be periodically updated to show trends in a particular timeframe, of the action schools should take in order to be successful on inspection. The data shows that common, replicated required actions are not unique by school size, location, starting age or inspector designation, and so would be considered applicable for all schools who desire to be graded at least 'good' in their next inspection. Someone utilising this software could give a statistically underpinned list of most desirable actions, as described by the inspectorate in any given timeframe.

It could be argued that more similarity should indicate more trust in the system. That similarity between reports implies a low incidence of independent interpretation of the Ofsted handbook and regulatory requirements, leading to more confidence that a 'good' judgement would be replicated by multiple inspectors over time. That the similarity implies a reduction in the influence of individual inspectors, lowering of bias (conscious or unconscious) and therefore demonstrates that the system is equitable and applied consistently. However, the similarity in the output of inspection (the report) and the similarity in the application of the handbook and policy during the process of inspection are separate variables. The report language is, arguably, where the school's unique context could be described and celebrated, rather than mediated into standardised content. This tension between similarity indicating consistency (and therefore implied validity), and similarity indicating homogenised or overly filtered content is considered across the findings.

Research question 2 (RQ2): Does the language used in Ofsted reports lead to subsequent improvements in outcomes?

Discourse analysis approaches, drawing on Fairclough (1989) and Van Dijk (2008), were applied to identify which parts of the reports were most unique. Using Turnitin and NVivo, the most unique part of the reports was found to be the 'areas for improvement' section, which summarises the larger sections of the report and focuses on specific improvements that schools are required to make before re-inspection. The areas for improvement use more direct language, condensed terminology, and a fractured stem/leaf structure of bullet points to give several directions with a limited word count. Given the uniqueness of the area for improvement sections, these were investigated as a stand-alone sub-group to explore whether specific uses of language correlated to changes in outcomes (RQ2).

The level of uniqueness of areas for improvement depended on the schools' grading. For schools graded less than 'good', these were on average more similar and full reports more bespoke; in 'good' or better schools, the areas for improvement were more bespoke and full reports more similar. This implies that the full report texts are used for clarification and exemplification, which is more detailed if schools are less than good and suggests the 'required actions' are more generic for those schools who have yet to be graded 'good'. The improvement actions are then more bespoke once schools have reached 'good', to specifically identify which sub-sections remain areas to improve within an already good school.

This tallies with my professional experience as a school leader and inspector, and makes logical sense given the role of the area for improvement section. However, the extent to which these parts of the report are more unique is significant: almost a third of areas for improvement are duplicated phrases for schools graded less than good. This again implies that there are common, generic actions all schools should complete to be graded 'good'. The implication of this finding suggests that the limited word count has a noteworthy impact on schools less than good, and that the merging of specific actions into broad overarching themes to comply with reporting restrictions causes reports for weaker schools to be less helpful than advice for schools already graded good.

Reports on inspections that started as Section 8 were found to focus more on generic descriptions and handbook terminology than specific curriculum elements, which could be due to the reduced time to plan curriculum activities within an amended inspection schedule. For example, if inspectors have not been able to observe sufficient teaching to be able to specify whether there are specific issues in, say, the teaching of mathematics or science, the areas for improvement may be worded as a need to achieve '*consistent teaching across subjects*' to be certain to cover the issue. However, on re-inspection, when this is followed up it will require that all subject teaching is consistent to demonstrate improvement in relation to this target. This indicates that those schools given generic 'areas to improve' from an initial Section 8 inspection have more work to do to improve.

Data showed that Section 8 deemed Section 5 inspection reports were more unique and had areas to improve that were more unique. This is estimated to be due to the differences in inspection protocols, and the lack of ability to report in detail on some sections because of time restrictions and late notice. It appears that the more generic improvement areas are written using language that is less similar to the corpus and cover areas that are broader or less specific. This group of less similar reports had no correlation to a difference in impact on outcomes for pupils.

Research question 3 (RQ3): Are there trends within the language used in Ofsted reports?

Those schools whose grading dropped during the inspection had more similar areas for improvement (31% similarity) and those schools whose grade improved had a much more unique areas for improvement (16% similarity). The language used in reports on weaker schools was very generic for those given a 'Requires Improvement' grading and significantly longer and more specific for schools graded 'Inadequate'. This follows the quality assurance principle of summarising weaknesses so that the audience does not think the school is worse than 'Requires Improvement', and the importance of highlighting any areas that are inadequate in detail. However, in practice, this suggests that those leaders in schools graded 'Requires Improvement' are tasked with completing actions that are overarching and less specific, and thereby a bigger workload.

Good schools, given more bespoke 'areas to improve', are less at liberty to make improvement or pedagogical choices, as the improvement actions are specific and detailed, with little to no room for interpretation. Re-inspection would require certain actions to have been undertaken, which do not always align with the content of the Ofsted handbook ‘outstanding’ criteria. For example, in the extracts below, the ‘good’ school is required to ‘extend vocabulary’, even though this is a pedagogical choice, which is explicitly described as not a part of inspection. Vocabulary is not mentioned in the inspection schedule, nor referenced in outcomes or teaching either directly or indirectly. The ‘Requires Improvement’ school is able to choose whether vocabulary is a factor in their actions to improve, and the ‘Inadequate’ school has to increase opportunities for extended writing in subjects and year groups and can decide whether vocabulary should be a part of that process.

> *Continue to increase the quality of teaching so that pupils make the best possible progress by… extending pupils’ use of interesting vocabulary when writing.*
>
> Report #10025179 (Good)

> *Improve pupils’ achievement by ensuring that all groups of pupils make good progress, as a result of teaching that…consistently provides the highest levels of challenge to enable pupils, especially the most able, to make good progress…*
>
> Report #10037015 (Requires Improvement)

> *Improve the quality of teaching so that all teaching is consistently good or better and raise outcomes for all pupils, including children in the early years, by… providing more opportunities to write independently in extended pieces of work and to apply their writing skills in other subjects…*
>
> Report #10025336 (Inadequate)

This relatively small finding has a substantial impact on the independence of school leaders to improve their schools and increase their Ofsted rating. If the areas for improvement are re-inspected and found to be lacking, then evaluations of leadership will be found to be poor, as leaders were given instructions on what aspects needed to improve and chose to reject these. However, if leaders do improve those aspects, and the next inspector cannot include or reference these improvements because they are not a part of the inspection schedule, then the efforts are not recognised. It is easier to re-inspect a requires improvement school with generic targets, as it is easier to find an action and evidence of improvement to a broad target than a specific one that might be impacted by cohort variance or context.

Research question 5 (RQ5): Within the area for improvement, are there any individual terms that are particularly effective or impactful?

To explore the extent to which the language of areas for improvement correlated with improvements to outcomes, reports on schools whose grading changed were tracked against outcomes data. First those schools whose grading dropped by more than one level, and who thereby were shown to have been given more generic areas for improvement were tracked against their combined KS2 Reading, Writing and Maths scores over the three years following that report. On average they improved by more than double the national average (+9.8% compared to national +4%). Upon re-inspection in 2019, one third remained at their less-than-good grading, and one third improved to be graded ‘good’. Roughly a third of these schools academized and were therefore not re-inspected. This suggests that the generic language was not a barrier to improvement for these schools. Whether it was the freedom to interpret improvement actions from generic areas for improvement, or whether it was simply that these schools had to improve a lot of things and leaders simply began an improvement process despite inspection is unclear. There is a correlation here, but no direct link other than a recognition that these schools had a public recognition of their weaknesses via inspection and made substantial efforts to improve.

For those schools who improved more than one grade, who had a more unique areas for improvement, the outcomes have a spiky profile, with four ‘super improvers’ skewing the data (progress measures of +17%, +29%, +14%, +22%). The super improvers all had entirely unique area for improvement (0% similarity) and comparatively unique reports (43%, 53%, 65%, 42% similarity). All other increased grade schools have a comparatively stable profile, with an average loss of -5% over the three years from just before inspection to the point of re-inspection for the rest of this group. Together, the average outcome evens out to +4.2% (national is +4%). This suggests that although these groups had a common level of similarity, there is no consistent positive correlation between an increase to a good grading, with a more bespoke area for improvement and a corresponding improvement in outcomes. All improved schools sustained their inspection grading for overall effectiveness and remained an average of 7% above national on combined reading, writing and maths scores, with a slight decline in attainment overall. This implies that those schools who improve their grading on inspection have no subsequent improvement patterns directly correlated to their areas for Improvement but sustained relatively strong performance.

Those schools who were graded inadequate had more unique language in the full report and areas for improvement and contained a greater emphasis on ‘actors’ (leaders, managers, governors, teachers). Some of the similarity arises from required content relating to governance and pupil premium reviews. The areas for improvement content is even more bespoke than percentages represent, therefore, as many of these actors were later taken out as ‘stop word’ phrases. Although their combined reading, writing and maths scores improved by +14% (national +4%), this is likely to be due to the very low performance that led to their inadequate grading than any specific terminology impact. The improvement language is heavily driven by the inadequate grading descriptors, and any improvement would be sufficient on re-inspection

The literature review had shown few projects correlating inspection practice to a measurable outcome on pupils (Chapter 2.5). Ehrens (2014) suggested links between inadequate schools and outcomes, and this was explored as part of the data (see Appendix 7 and Chapter 4.5). To enable this research to investigate the impact on outcomes of the area for improvement, previously identified as the most powerful and unique section of the report, a single, measurable curriculum area was isolated. From a discourse and word frequency analysis of the improvement section, key terms were identified as more frequent and more regularly duplicated (progress, skills, consistent) and some of these related to measurable outcomes for pupils in terms of nationally recognised data (reading, writing, phonics). There were also frequent key terms that related to curriculum skills and knowledge (calculation, spelling, reasoning). Of these, mathematical language was the most frequently cited within the corpus (1028 references, compared to 982 for writing, 641 for reading). Mathematics outcomes for pupils were publicly available as well as national data. Data was investigated for both progress (improvement) and attainment (final outcomes) for Key Stage 2 pupils and tracked back to the area for improvement given during the inspection. This iterative refinement of RQ2 was critical to be able to follow a group of reports to an outcome for pupils, and followed mixed methods approaches outside of traditional discourse analysis techniques (Creswell, 2014).

This finding is interesting for schools who have been given a good judgement. If this data is accepted, it would suggest that schools who are graded Inadequate should closely follow the requested actions from the area for improvement section, which are detailed and bespoke. Equally, for schools judged ‘good’ the advice and required actions are specific and tailored closely to that particular school and should be considered representative and useful. However, for requires improvement schools, the areas for improvement are generally overly broad and non-specific, and do not give specific or bespoke advice, and so can be considered less supportive of successful school improvement.

Refined research question 2) What is the impact of an ‘area for improvement’ relating to mathematics on subsequent Mathematics outcomes?

The analysis showed that those schools given mathematics as an area to improve subsequently had progress improve seven times faster than the national average. Attainment, in the form of average scaled scores, improved marginally better than the national average. This correlation between schools who are required to improve using mathematics terminology and greater improvement in children’s mathematics scores than similar schools in the same period suggests that being given a mathematics directive to improve can lead to improved performance in that area. This is in opposition to many qualitative projects (Coffield, 2012; Altrichter and Kemethofer, 2015) which cite the detrimental impact on staff workload, morale, and independence from the pressure of inspection as being more influential than the directives themselves, and only marginal impact on pupil outcomes. Those projects describe marginal influence on outcomes in individual schools, but significant influence on school cultures as the prevailing outcome of inspection. This project, looking at the data quantitatively, using national averages and over several years, indicates that the act of reporting a requirement target for mathematics does appear to have a positive impact on outcomes in mathematics.

The link is implied but cannot be guaranteed due to the myriad of other variables at play within the system at the individual level, but this macro-level data shows that those schools who were told to improve mathematics, when considered as a group, did so at a rate significantly faster than all schools for both progress and attainment. Earlier projects, such as those of Wilcox and Grey (1996) suggested the impact of inspection was neutral overall, based on HMI-led inspections, and Rosenthal (2004) indicated a slight negative impact, so this would be a new finding for the sector. This is possibly due to the ‘cumulative mean’ effect of earlier research measuring impact across multiple variables (reading, writing, maths, combined) rather than tracking one individual area for improvement focus to its associated data set, which has only recently been possible on such a large scale.

A clear trend was found in terms of mathematics. No matter how the term was included, whether as a very specific target or general requirement, if a school was told to improve mathematics, the data shows that the schools’ results in that area improved faster than national averages. Many areas for improvement that included mathematics were unique and a large proportion were from previously good or outstanding schools. This suggests that even schools already performing well improved their progress and outcomes in mathematics.

To drill down into the mathematics finding, the exact phrasing of the areas for improvement were further investigated to see if there were any key terms that correlated with improved outcomes using Big Data strategies and established discourse analysis principles (Van Dijk, 2001) such as isolating single words. The words ‘problem solving’, ‘reasoning’, and ‘calculation’ correlated with measurable improvement in outcomes. These were investigated individually and in sequence for statistical significance. Other terms, such as ‘number’ and ‘accuracy’, were less frequent within the corpus but culturally represent specific improvement activity (teachers working towards a single, measurable curriculum objective) which could be replicated in multiple settings. To be confident of correlation for these specific terms a larger data set would be required, as their occurrences were significantly fewer in this corpus.

Where inspectors had used terminology that was not only mathematics-specific, but that also references a particular set of learning objectives, assessment criteria or section of the Standard Assessment Tests (SATS), the improvement was significantly above national average, and where terms were used in combination, the cumulative improvement was increased (see appendix 15, 16). The power of these terms may not be easily recognised by non-specialists as a curriculum element, but each relates to a specific sub-set of mathematical knowledge and skills that staff in schools can identify and teach towards, leading to concrete improvements in testing scores.

Checks on this hypothesis that specific mathematics terminology correlated with improved outcomes in maths, different mathematical and non-mathematical terms were investigated. All mathematical terms correlated to increased progress or attainment, and those without those terms did not show subsequent trends. Terms with equitable frequency were checked such as ‘pupils’ or ‘teachers’ and results reflected the mean, rather than any positive trend. Terms such as ‘presentation’ or ‘challenge’ had varied results, including some negative associations, and those that did not include specific mathematical terms were found to be not significant. Non-mathematical specific terms below an overall improvement area stem for mathematics did reflect improvement generally. This supports the finding as significant.

Most schools that had been given mathematics as an area to improve would be expected to focus training and monitoring activity on mathematics for the years following inspection, but whether this comes at the expense of other subjects and causes an overall decline in other or combined scores might potentially be why previous research, using combined scores in reading, writing and maths, or average overall attainment, has shown less correlation between inspection and impact on outcomes. From experience, many schools who focus on one subject often see associated declines in others, as staff capacity is limited, especially in smaller schools.

This evidence is limited by there being a very small number of schools that had nil returns for some Standard Assessment Test (SATS) data during the period and therefore could not be included, and improvement in progress or attainment over time could not be comparably calculated. Therefore, this sample excludes a small proportion of schools who did have maths as an area to improve, and means the set is skewed away from very small settings, with outcomes data that may have been withheld for privacy reasons.

The implications of this finding are that where inspectors have used mathematical terms associated with the Standard Assessment Tests within their areas for improvement, those schools on average, will improve mathematics scores to a significantly better degree than the national average. This has implications for equity and consistency of inspection practice, for the actions taken by schools, and in the writing of areas for improvement. Given that the project has also shown that those schools graded ‘requires improvement’ are given generic, non-specific terms, then this implies they will be less likely to improve their mathematics outcomes. Given that reports on those schools graded ‘good’ and those from Section 5 inspections are more likely to state specific curriculum elements, these schools are more likely to improve mathematics outcomes. These discrepancies over time could lead to some schools benefitting far more from inspection than others

Chapter Six

Conclusion

From this research it can be said that Ofsted reports are not unique. A large proportion of required content, relating to structural norms, inclusion of mandatory phrases and terms mean that more than half of an average primary school Ofsted report is duplicated content, not unique to that setting (chapter 4.2, appendix 3)

The implications of these findings are that although full Ofsted reports are highly similar and represent an increasingly iterative reporting process with limited bespoke content, this has very little influence on the performance of schools after inspection. Where the report does appear to have power and influence is within the areas for improvement section, (chapter 4.8) where it is found that relatively minor choices of bespoke language and phraseology closely correlate with an improvement in outcomes as measured by SAT scores. Where areas for improvement include terms directly relating to SATs in mathematics, improvement is substantially greater and more rapid than the national average (chapter 4.9, appendix 15 and 16). If the inspectorate chooses to recognise this finding, it could validate whether there is equivalent correlation between English or other curriculum areas and subsequent improvements and using this knowledge could significantly improve the impact of reporting on improving outcomes in schools.

It would appear that the level of similarity is not the overriding factor, but rather the use of specific terminology that enables leaders to pinpoint specific areas of the curriculum and particularly those areas that are tested, in order to improve outcomes for pupils in that element (figure 13). This does appear to suggest a more audit-centric rather than descriptive/bespoke reporting approach is more effective, which is in opposition to what is considered current Ofsted ideology (Ofsted Handbook 2021), where data is being removed from the inspection process as much as possible, and the ‘outcomes’ judgement has been replaced with ‘quality of education’ – a much more general and descriptive grouping.

Implications

- That school leaders should focus on the areas for improvement sections in not only their own setting, but similar settings inspected recently to identify where they could improve so as to perform better on re-inspection
- That the terminology of areas for improvement should be used carefully by inspectors, as their power and weighting are disproportionate compared to the rest of the report

- That those being inspected should push for more specific terminology within their areas for improvement where possible, rather than generic areas.
- That the inspectorate should consider replicating the model to look for impact on reading, writing, phonics, etc. and in the secondary sector to gather concrete data on the impact of the inspection process.

Further Lines of Enquiry

Although not investigated within this project, further research could be undertaken to compare schools who have significantly similar reports, to see the extent of similarity visible from evidence within school, exploring whether this linguistic shorthand has skewed evaluation of their settings, or if the process slots unique practice into generic ranking brackets that have identical overarching descriptors, yet bespoke embodiment within school. For example, since Ofsted began describing subject-level scrutiny as a ‘deep dive’ in 2019, many schools changed their policies and documentation to include activities called ‘deep dives’ generating evidence in preparation for inspection. The term ‘deep dive’ was introduced by the inspectorate to describe one of their evidence collecting methods, and now in this simple two-word phrase, describes a range of education quality assurance processes covering observation, book scrutiny, staff meetings, planning and assessment scrutiny that is understood and agreed across settings and has become accepted school terminology. The introduction of this term added workload, changed quality assurance processes and the nature of evidence gathered by schools. Due to this influence of the language of inspection, relatively minor changes to the terminology within reports have an exponentially bigger impact within schools, as the language and associated actions are so heavily duplicated and due to this, carry power to change practice in schools. There were no new ‘fashionable’ language trends within this corpus, possibly because of the deliberate attempt to capture a time period within a single Ofsted handbook variation. It would be an interesting project to track reports across a change in handbook, to note the proliferation of new terms from the handbook to the report, and to actions required.

In terms of language trends, if inclusion of the Standard Assessment Test terminology is not subsumed into the ‘voice of the inspectorate’ and does not become standard inspection lexicon, then there could be a significant and visible discrepancy between the possibility of reports to impact on improving outcomes in mathematics. If the pattern is consistent across Areas for Improvement relating to English standard tests and terms (which are not investigated in this project) then the pattern could be considered even more imperative to integrate within the ‘voice of the inspectorate’.

The method of using Ofsted reports as evidence of the work of the inspectorate could be a rich source of further data. Since the project, inspectors have begun to work within regional areas, so that the ability to scrutinise linguistic differences between regions is now possible. Equally, HMI now manage teams of regional Ofsted inspectors, so the opportunity for language duplication between trainees and established inspectors within a smaller zone of proximity is increased. Discourse analysis literature suggests that a smaller, geographically dense team would increase the likelihood of adopted shorthand terms (Van Dijk, 2008). There are also new data sets published by the Department for Education on the regional performance of schools, so a closer correlation between the work of the inspectorate in a particular geographical area, and its impact on school outcomes could be analysed in future projects.

The project also shows that requires improvement schools have different advice and required action profiles than other settings. The implication of this is that it would be easier for a requires improvement school to demonstrate improvement and gain a good grading on inspection than a previously good school. Further study of grade profile changes compared to area for improvement phrases would need to be done to investigate this potential finding.

For inadequate schools, the language of inadequate reports, although more bespoke and more clearly describing context and characteristics, has less of an impact on improvement than the overall judgement itself, which often triggers local authority, DFE or Board-level intervention as well as the pressure of imminent re-inspection. It would be difficult to identify any trends from this very small sub-group’s terminology that could be generalised sufficiently to form any causative links. Further investigation into inadequate graded schools, and the practices deployed to help them improve would be an interesting project, to see how tied to the Ofsted language of improvement and action plans were, and if any correlation over time and a larger sample group could be identified.

Applications to practice

Since undertaking this research, several parts of the method of data analysis have proven helpful in my professional practice in a leadership role over a large group of schools. I have used the discourse analysis processes frequently, for example to code inspection reports for our schools, specifically focusing on the areas for improvement as the most critical and influential part of the documents. I have been able to separate and code our own reports and compare them to the reports of our direct competitors and to all reports across the sector. This has saved hours of workload, scrutinising, and synthesising the information across a large trust to identify emerging trends and support funding, training, and improvement strategies. To date, we have been able to identify shared training for specific curriculum elements that were beginning to trend as an area for improvement, and these are now reported on as strengths. We were able to better articulate value for money, when that became a criticism in multiple reports, and have generated templates and support documents to better prepare our schools for emerging themes.

It has shown us whether the trends within our own schools reflect the sector at large, or if there are trends in the advice given to others that we can learn lessons from before our own schools are inspected. I have used it to synthesise reports for the executive board, so I can analyse how we compared to the last term or year, scrutinised our reports and areas for improvement for sentiment, so I can see if our own internal quality assurance themes match Ofsted's opinion of our schools, and have even used it on internal reports, showing themes across our consultant or governor reports, so I can balance the subject or leadership focus of our school improvement visits and challenge.

As the highest performing 'large provider' according to Ofsted, with more than 50 schools and over a hundred residential settings as well as multiple fostering agencies, this ability to synthesise and quickly and efficiently scan and summarise multiple documents has proven invaluable. The tools I have designed and used within this project have given us rapid and clear access to the detail of advice and guidance across the group while also indicating the distinctiveness of individual reports. Together, the tools are helping us to shape and refine support, challenge, and accountability across hundreds of staff and schools. The identification of specific curriculum areas in need of improvement from multiple reporting sources has meant that we can tackle and refine those elements before inspection. Where schools have performed less well on inspection in some areas for our competitors, we can spotlight that particular regulation and prepare our leaders to better articulate our compliance and strengths. The possibility for senior educational leaders to repurpose research tools in this way is an important outcome from this project and is as much a critical finding as the data itself. I have already been disseminating the use of NVivo and Turnitin for quality assurance, monitoring, and data mining to a range of external partners.

The initial purpose of the project was to investigate the extent to which the perception of school leaders about how standardised the reporting process had become, and therefore whether the inspection process was able to capture their unique practice and setting. This was illustrated by school leaders being unable to recognise whether a page from a report was one written about their school. At a recent head's meeting I showed some anonymised front-page bullet points to a group of heads of 'good' schools and almost all failed to identify the one that was from their school. This further showed me that it is not the detail of the reports which heads use to drive improvement and actions, but the areas for improvement sections (which almost all correctly identified to their own school) and that, given its power and influence, the language used in this part of the report is of critical importance if it is to lead to improved outcomes for children.

As of 2019 Ofsted were releasing research papers and blog posts (nuance studies, 2019) to support their methodology and confirm that multiple inspectors would reach the same conclusion on an inspection visit, which further underlined the standardisation and performance of the inspection regime. A paper on the reliability of short (section 8) inspections, released as part of this series, agreed with the findings of Chapter 4.4 relating to the similarity between section 5 and deemed inspections overall. The tension between similarity meaning consistent and equitable, and similarity meaning 'watered down' and generic remains. Although trust in the reliability of evaluation is important, the accuracy of reporting as a measurement compared to the power of reporting as a tool for improvement remains a critical tension. If reporting has a similarly powerful impact on other areas of the curriculum, then inspection is a much more valuable tool for schools as a mechanism for improvement, which would balance out concerns over consistency of judgements between inspectors and over time.

Post-script: Coming up to date

In the time following this project, inspections were briefly paused, then moved to remote, and are now being re-introduced as face-to-face events (Ofsted's plans 2021). The pandemic has changed the way data is used to triangulate, as for 2020 and 2021 assessment in primary schools was unable to be undertaken under normal protocols due to the substantial absences and impact of the lockdown measures. This has meant that there is a much weaker link between reports and schools, and some reports are now drastically out of date. Due to restrictions on physical visits to schools, Ofsted extended the normal periods between inspections and as of the end of 2021, some are being inspected on a traditional cycle, others on a 'catch up' timeline. Section 8 inspections continue to be used. The pandemic also forced the handbook to be amended to reflect some expected activities such as trips, visits, or work experience, which are limited or unavailable, and some schools who retain partial 'remote learning' offers for pupils due to medical or physical reasons. As such, the ability to relate reports from March 2020 onwards to those prior, is limited, as the schools and settings they describe are very different from their previous incarnations. As results are skewed by factors largely outside of school's control during this period (access to technology, home situations, etc.) any correlations between inspection outcomes and children’s outcomes for this period are very tenuous. To reflect this, inspections undertaken in Autumn 2021 are focussing much more on school’s ability to plan, sequence and link learning for pupils, with areas for improvement reflecting curriculum leadership far more than progress or outcomes. How this will impact future school improvement plans, which are now more responsive to the short-term fluctuations in attendance, medical and government advice rather than inspection pressures is yet to be seen.

Appendix

An appendix has been included to exemplify some of the core terms and some of the more detailed data for those who may be interested in replicating or taking up some of the abandoned trails that this iterative process identified.

Appendix 1 shows the initial spreadsheet, designed to collate the variables associated with each report, and Appendix 2 detail of the corpus at each stage of iteration, with the proportionate team sizes, gender and designation of lead inspectors and judgement profiles.

Appendix 3 is a graph of the distribution of similarity for the full reports, showing the spread of number of reports at different similarities is a rough bell curve, and Appendix 4 shows the distribution of largest single source duplication for the full reports, showing that most reports had the largest duplication from one source of around 15%. This is discussed in chapter 4.2.

Appendix 5 is a table showing the associated variables for different lead inspector genders and designations, and their variance from the corpus as a whole (Chapter 4.7)

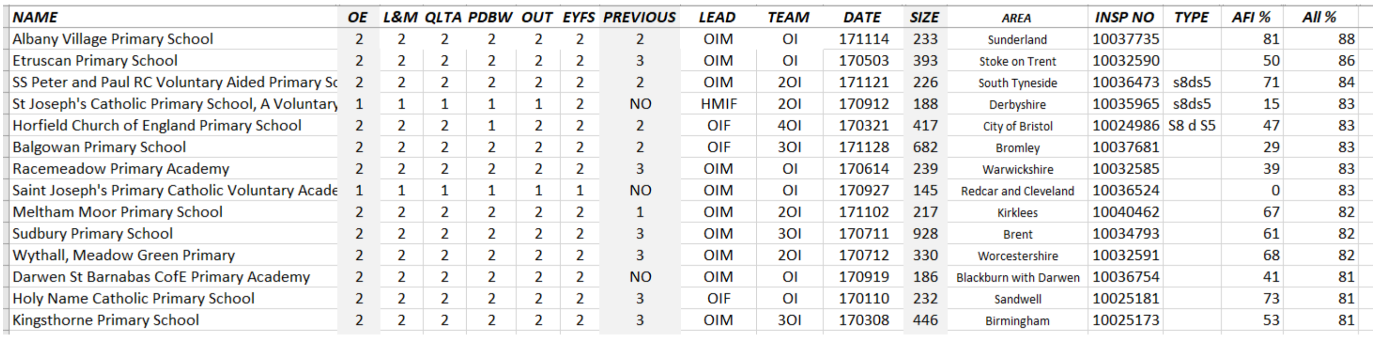
Appendix 6 describes the word length of those reports graded inadequate within the corpus, and Appendix 7 is the detailed grid showing the performance of all schools in the corpus graded inadequate, investigated in detail in Chapter 4.6 in case correlating findings to existing research focussing on this group of schools became apparent.

Appendix 8 shows the number of references coded by NVivo at each ‘sentiment’ level, and Appendix 9 the top 50 words under each sentiment (Chapter 4.1).

Appendix 10 and Appendix 11 show the word frequency searches using NVivo and their ranking, and those words grouped into themes using professional insight, in case trends could be ascertained. Appendix 12 and 13 show change in frequency rankings for section 5 compared to section 8 (Chapter 4.4)

Appendices 14, 15 and 16 show the details of the T-Test results, comparison groups and the associated values. As these are prolific, only those showing significance are listed for exemplification. Other combinations can be assumed to be not showing as significant in this data set.

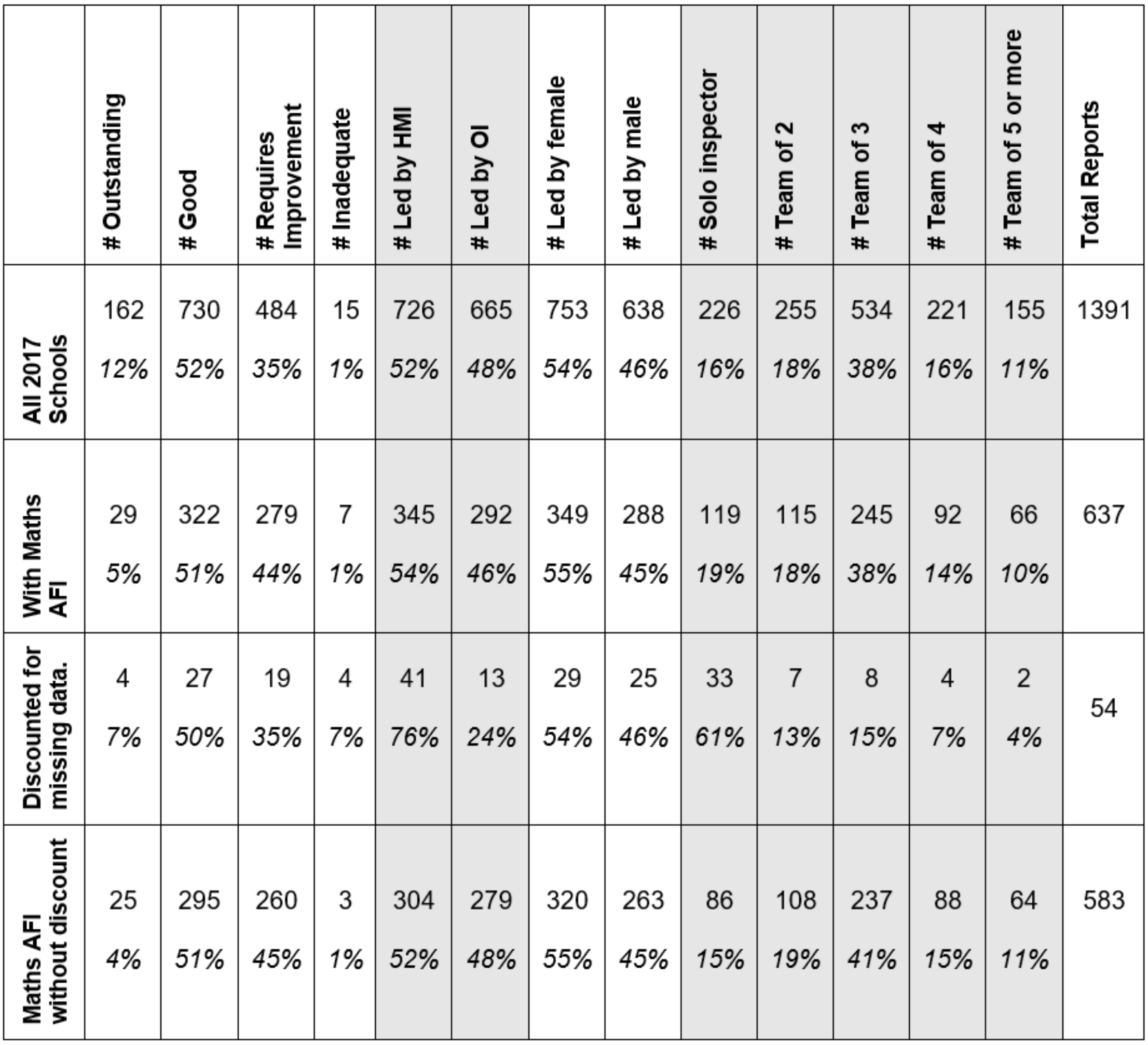
Appendix 1 - Exemplar showing spreadsheet layout for initial sample scrutiny.



NAME	OE	L&M	QLTA	PDBW	OUT	EYFS	PREVIOUS	LEAD	TEAM	DATE	SIZE	AREA	INSP NO	TYPE	AFI %	All %
Albany Village Primary School	2	2	2	2	2	2	2	OIM	OI	171114	233	Sunderland	10037735		81	88
Etruscan Primary School	2	2	2	2	2	2	3	OIM	OI	170503	393	Stoke on Trent	10032590		50	86
SS Peter and Paul RC Voluntary Aided Primary Sc	2	2	2	2	2	2	2	OIM	2OI	171121	226	South Tyneside	10036473	s8ds5	71	84
St Joseph's Catholic Primary School, A Voluntary	1	1	1	1	1	2	NO	HMIF	2OI	170912	188	Derbyshire	10035965	s8ds5	15	83
Horfield Church of England Primary School	2	2	2	1	2	2	2	OIF	4OI	170321	417	City of Bristol	10024986	S8 d S5	47	83
Balgowan Primary School	2	2	2	2	2	2	2	OIF	3OI	171128	682	Bromley	10037681		29	83
Racemeadow Primary Academy	2	2	2	2	2	2	3	OIM	OI	170614	239	Warwickshire	10032585		39	83
Saint Joseph's Primary Catholic Voluntary Acade	1	1	1	1	1	1	NO	OIM	OI	170927	145	Redcar and Cleveland	10036524		0	83
Meltham Moor Primary School	2	2	2	2	2	2	1	OIM	2OI	171102	217	Kirklees	10040462		67	82
Sudbury Primary School	2	2	2	2	2	2	3	OIM	3OI	170711	928	Brent	10034793		61	82
Wythall, Meadow Green Primary	2	2	2	2	2	2	3	OIM	2OI	170712	330	Worcestershire	10032591		68	82
Darwen St Barnabas CofE Primary Academy	2	2	2	2	2	2	NO	OIM	OI	170919	186	Blackburn with Darwen	10036754		41	81
Holy Name Catholic Primary School	2	2	2	2	2	2	3	OIF	OI	170110	232	Sandwell	10025181		73	81
Kingsthorne Primary School	2	2	2	2	2	2	3	OIM	3OI	170308	446	Birmingham	10025173		53	81

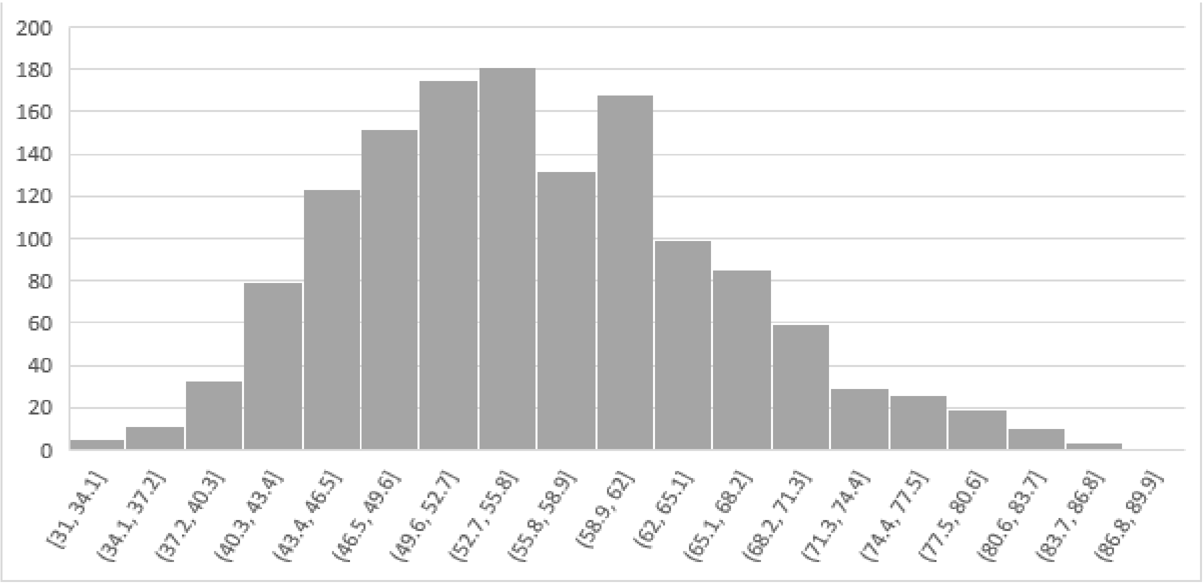
Columns include visible attributes, similarity scores were added and later, progress and attainment data for mathematics. Titles are 'OE – overall evaluation, L&M leadership and management, QLTA – quality of learning teaching and assessment, PDBW, personal development behaviour and wellbeing, OUT – outcomes, EYFS - early years foundation stage, 'previous' – last overall evaluation judgement (NO = not previously inspected), 'lead' – whether an OI - Ofsted Inspector (M- male, F-female) or HMI (her majesty's inspector), number of Ofsted inspectors in the team, the date of inspection, size of the school by number of pupils, geographical area, inspection number, type of inspection (marking whether a section 8 deemed section 5) and the similarity percentage of that report using Turnitin for just the area for improvement section or all sections of the report (ALL).

Appendix 2 – Broad features of the corpus at each iteration**.**



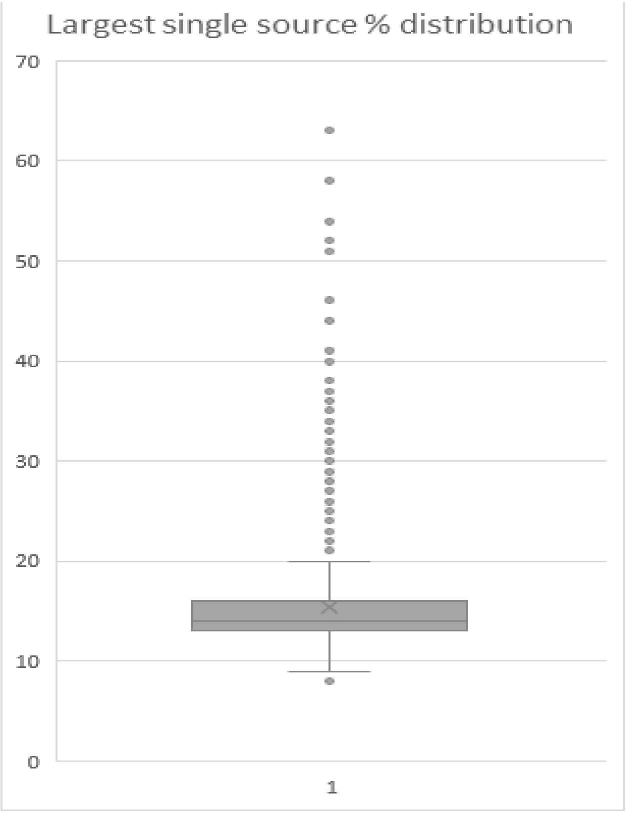
	# Outstanding	# Good	# Requires Improvement	# Inadequate	# Led by HMI	# Led by OI	# Led by female	# Led by male	# Solo inspector	# Team of 2	# Team of 3	# Team of 4	# Team of 5 or more	Total Reports
All 2017 Schools	162 *12%*	730 *52%*	484 *35%*	15 *1%*	726 *52%*	665 *48%*	753 *54%*	638 *46%*	226 *16%*	255 *18%*	534 *38%*	221 *16%*	155 *11%*	1391
With Maths AFI	29 *5%*	322 *51%*	279 *44%*	7 *1%*	345 *54%*	292 *46%*	349 *55%*	288 *45%*	119 *19%*	115 *18%*	245 *38%*	92 *14%*	66 *10%*	637
Discounted for missing data.	4 *7%*	27 *50%*	19 *35%*	4 *7%*	41 *76%*	13 *24%*	29 *54%*	25 *46%*	33 *61%*	7 *13%*	8 *15%*	4 *7%*	2 *4%*	54
Maths AFI without discount	25 *4%*	295 *51%*	260 *45%*	3 *1%*	304 *52%*	279 *48%*	320 *55%*	263 *45%*	86 *15%*	108 *19%*	237 *41%*	88 *15%*	64 *11%*	583

Appendix 3 Similarity frequency in groups of 3% from 31% (least duplicated) to 88% (most duplicated) showing the spread of similarity across the corpus

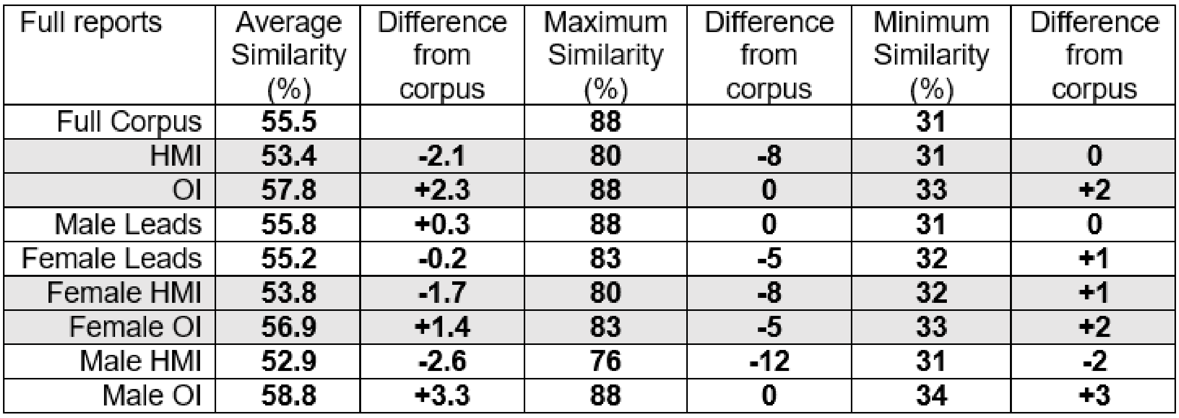


Appendix 4

Showing the proportion of largest single source distribution, (largest duplication from one other report). Shows data skewed towards most reports having around 15% from one other single report within the corpus, with some much higher duplications from a single other report, and no report being fully unique.



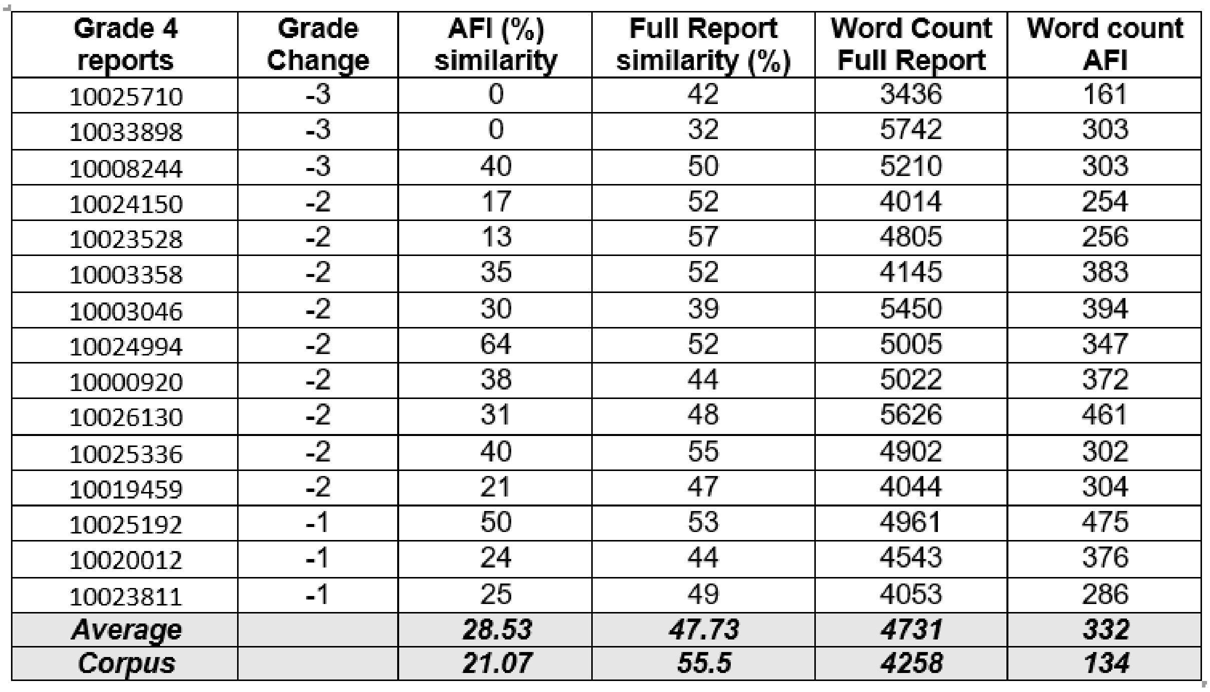
Appendix 5 Lead Inspector similarity trends



Full reports	Average Similarity (%)	Difference from corpus	Maximum Similarity (%)	Difference from corpus	Minimum Similarity (%)	Difference from corpus
Full Corpus	**55.5**		**88**		**31**	
HMI	**53.4**	**-2.1**	**80**	**-8**	**31**	**0**
OI	**57.8**	**+2.3**	**88**	**0**	**33**	**+2**
Male Leads	**55.8**	**+0.3**	**88**	**0**	**31**	**0**
Female Leads	**55.2**	**-0.2**	**83**	**-5**	**32**	**+1**
Female HMI	**53.8**	**-1.7**	**80**	**-8**	**32**	**+1**
Female OI	**56.9**	**+1.4**	**83**	**-5**	**33**	**+2**
Male HMI	**52.9**	**-2.6**	**76**	**-12**	**31**	**-2**
Male OI	**58.8**	**+3.3**	**88**	**0**	**34**	**+3**

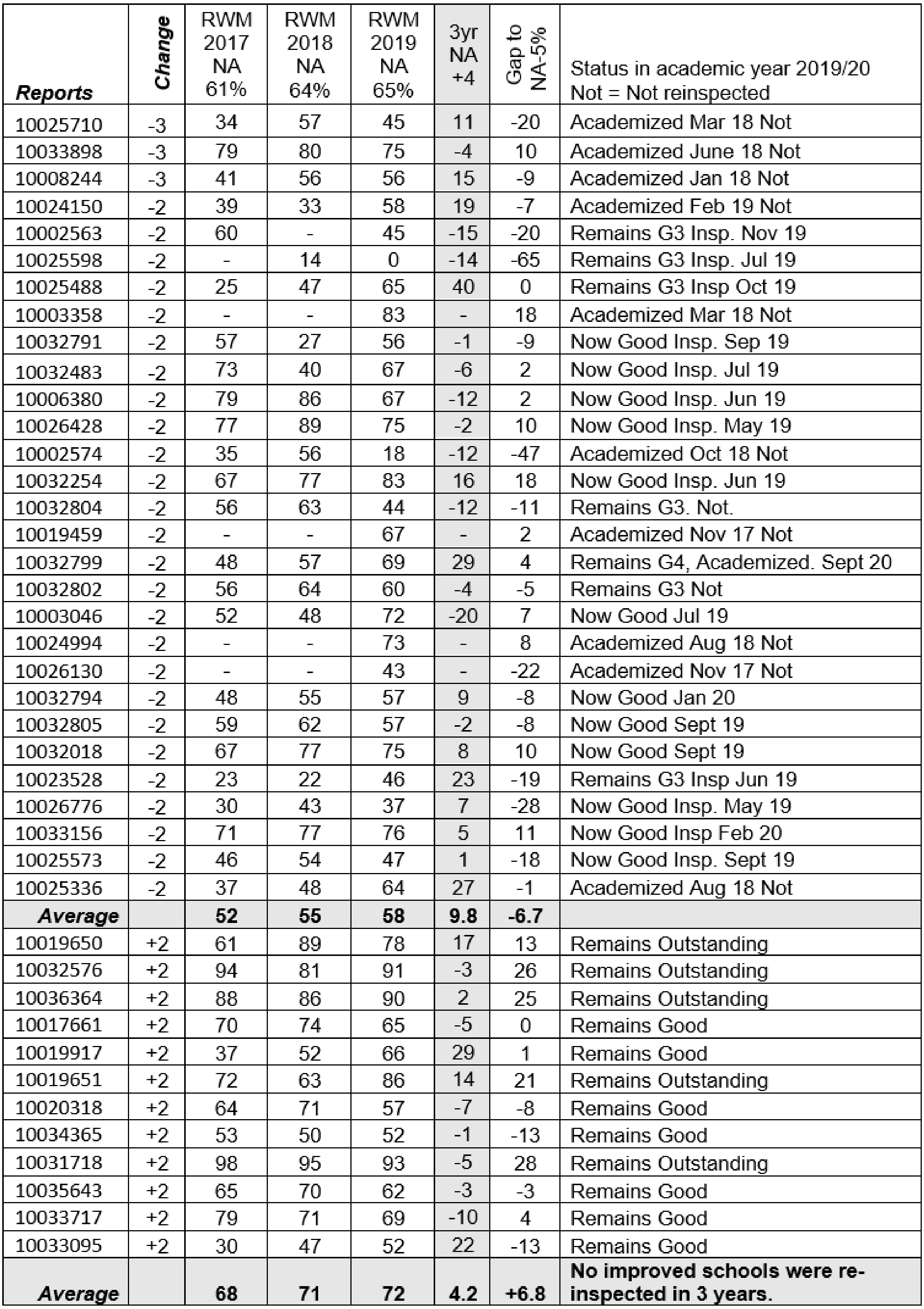
Showing that there is negligible difference in the similarity profiles of different lead inspector types for the full reports.

Appendix 6 Inadequate report comparative similarity and length



Grade 4 reports	Grade Change	AFI (%) similarity	Full Report similarity (%)	Word Count Full Report	Word count AFI
10025710	-3	0	42	3436	161
10033898	-3	0	32	5742	303
10008244	-3	40	50	5210	303
10024150	-2	17	52	4014	254
10023528	-2	13	57	4805	256
10003358	-2	35	52	4145	383
10003046	-2	30	39	5450	394
10024994	-2	64	52	5005	347
10000920	-2	38	44	5022	372
10026130	-2	31	48	5626	461
10025336	-2	40	55	4902	302
10019459	-2	21	47	4044	304
10025192	-1	50	53	4961	475
10020012	-1	24	44	4543	376
10023811	-1	25	49	4053	286
Average		***28.53***	***47.73***	***4731***	***332***
Corpus		***21.07***	***55.5***	***4258***	***134***

Appendix 7 Performance of those schools graded Inadequate from the corpus

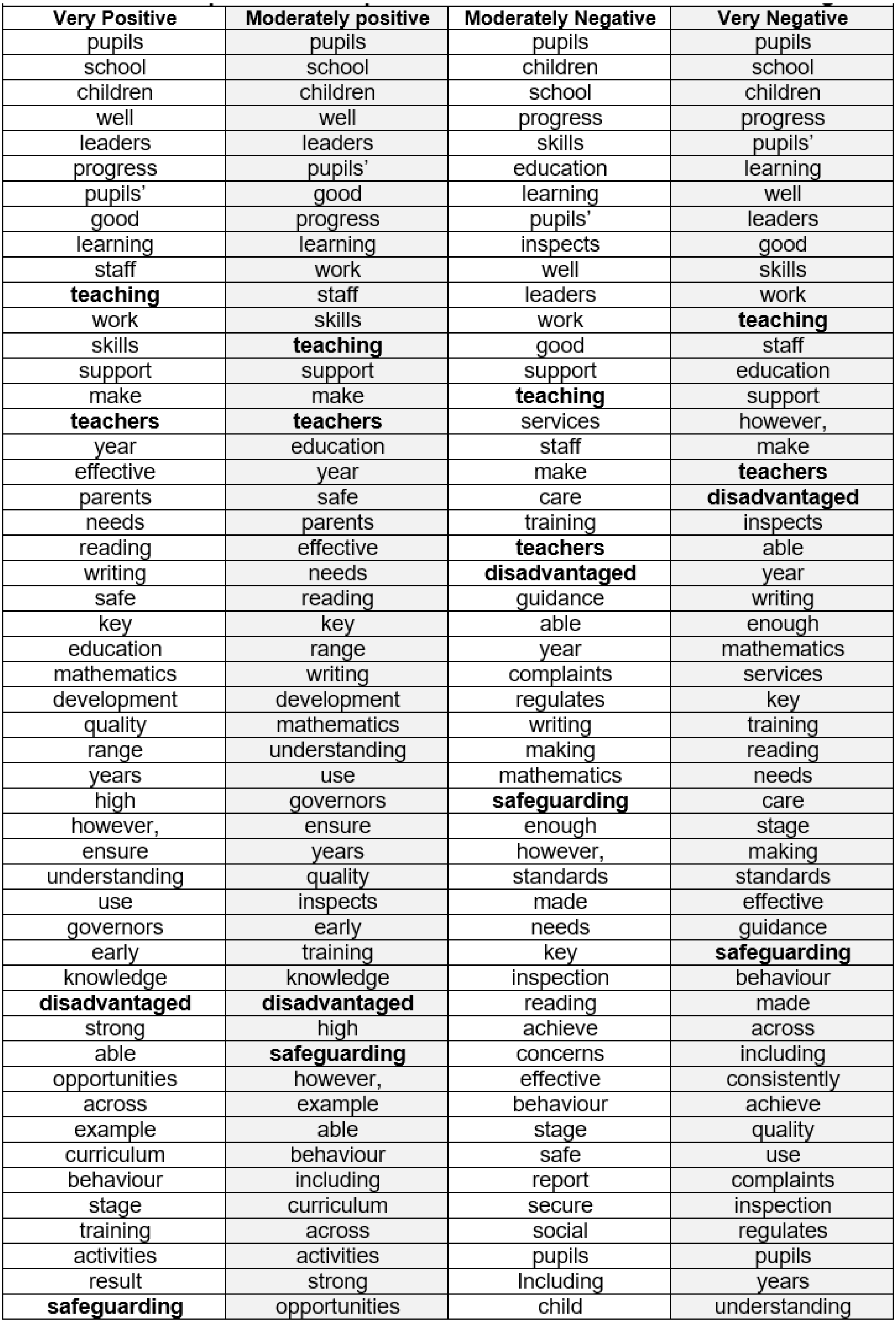


Reports	***Change***	RWM 2017 NA 61%	RWM 2018 NA 64%	RWM 2019 NA 65%	3yr NA +4	Gap to NA-5%	Status in academic year 2019/20 Not = Not reinspected
10025710	-3	34	57	45	11	-20	Academized Mar 18 Not
10033898	-3	79	80	75	-4	10	Academized June 18 Not
10008244	-3	41	56	56	15	-9	Academized Jan 18 Not
10024150	-2	39	33	58	19	-7	Academized Feb 19 Not
10002563	-2	60	-	45	-15	-20	Remains G3 Insp. Nov 19
10025598	-2	-	14	0	-14	-65	Remains G3 Insp. Jul 19
10025488	-2	25	47	65	40	0	Remains G3 Insp Oct 19
10003358	-2	-	-	83	-	18	Academized Mar 18 Not
10032791	-2	57	27	56	-1	-9	Now Good Insp. Sep 19
10032483	-2	73	40	67	-6	2	Now Good Insp. Jul 19
10006380	-2	79	86	67	-12	2	Now Good Insp. Jun 19
10026428	-2	77	89	75	-2	10	Now Good Insp. May 19
10002574	-2	35	56	18	-12	-47	Academized Oct 18 Not
10032254	-2	67	77	83	16	18	Now Good Insp. Jun 19
10032804	-2	56	63	44	-12	-11	Remains G3. Not.
10019459	-2	-	-	67	-	2	Academized Nov 17 Not
10032799	-2	48	57	69	29	4	Remains G4, Academized. Sept 20
10032802	-2	56	64	60	-4	-5	Remains G3 Not
10003046	-2	52	48	72	-20	7	Now Good Jul 19
10024994	-2	-	-	73	-	8	Academized Aug 18 Not
10026130	-2	-	-	43	-	-22	Academized Nov 17 Not
10032794	-2	48	55	57	9	-8	Now Good Jan 20
10032805	-2	59	62	57	-2	-8	Now Good Sept 19
10032018	-2	67	77	75	8	10	Now Good Sept 19
10023528	-2	23	22	46	23	-19	Remains G3 Insp Jun 19
10026776	-2	30	43	37	7	-28	Now Good Insp. May 19
10033156	-2	71	77	76	5	11	Now Good Insp Feb 20
10025573	-2	46	54	47	1	-18	Now Good Insp. Sept 19
10025336	-2	37	48	64	27	-1	Academized Aug 18 Not
Average		**52**	**55**	**58**	**9.8**	**-6.7**	
10019650	+2	61	89	78	17	13	Remains Outstanding
10032576	+2	94	81	91	-3	26	Remains Outstanding
10036364	+2	88	86	90	2	25	Remains Outstanding
10017661	+2	70	74	65	-5	0	Remains Good
10019917	+2	37	52	66	29	1	Remains Good
10019651	+2	72	63	86	14	21	Remains Outstanding
10020318	+2	64	71	57	-7	-8	Remains Good
10034365	+2	53	50	52	-1	-13	Remains Good
10031718	+2	98	95	93	-5	28	Remains Outstanding
10035643	+2	65	70	62	-3	-3	Remains Good
10033717	+2	79	71	69	-10	4	Remains Good
10033095	+2	30	47	52	22	-13	Remains Good
Average		**68**	**71**	**72**	**4.2**	**+6.8**	**No improved schools were re-inspected in 3 years.**

Appendix 8 Table showing the number of references coded at each sentiment (positive or negative) by NVivo for the full reports.

	Total # reports	**Total References**	**Proportion of references**
Total Positive	1391	65925	
Very Positive	1390	15368	23%
Moderately Positive	1391	50557	77%
Total Negative	1391	22856	
Moderately Negative	1391	16233	71%
Very Negative	1346	6623	29%

Appendix 9 50 most frequent words within each sentiment coding



Very Positive	Moderately positive	Moderately Negative	Very Negative
pupils	pupils	pupils	pupils
school	school	children	school
children	children	school	children
well	well	progress	progress
leaders	leaders	skills	pupils’
progress	pupils’	education	learning
pupils’	good	learning	well
good	progress	pupils’	leaders
learning	learning	inspects	good
staff	work	well	skills
teaching	staff	leaders	work
work	skills	work	**teaching**
skills	**teaching**	good	staff
support	support	support	education
make	make	**teaching**	support
teachers	**teachers**	services	however,
year	education	staff	make
effective	year	make	**teachers**
parents	safe	care	**disadvantaged**
needs	parents	training	inspects
reading	effective	**teachers**	able
writing	needs	**disadvantaged**	year
safe	reading	guidance	writing
key	key	able	enough
education	range	year	mathematics
mathematics	writing	complaints	services
development	development	regulates	key
quality	mathematics	writing	training
range	understanding	making	reading
years	use	mathematics	needs
high	governors	**safeguarding**	care
however,	ensure	enough	stage
ensure	years	however,	making
understanding	quality	standards	standards
use	inspects	made	effective
governors	early	needs	guidance
early	training	key	**safeguarding**
knowledge	knowledge	inspection	behaviour
disadvantaged	**disadvantaged**	reading	made
strong	high	achieve	across
able	**safeguarding**	concerns	including
opportunities	however,	effective	consistently
across	example	behaviour	achieve
example	able	stage	quality
curriculum	behaviour	safe	use
behaviour	including	report	complaints
stage	curriculum	secure	inspection
training	across	social	regulates
activities	activities	pupils	pupils
result	strong	Including	years
safeguarding	opportunities	child	understanding

Appendix 10 Showing the 200 most frequent words from the full corpus using NVivo

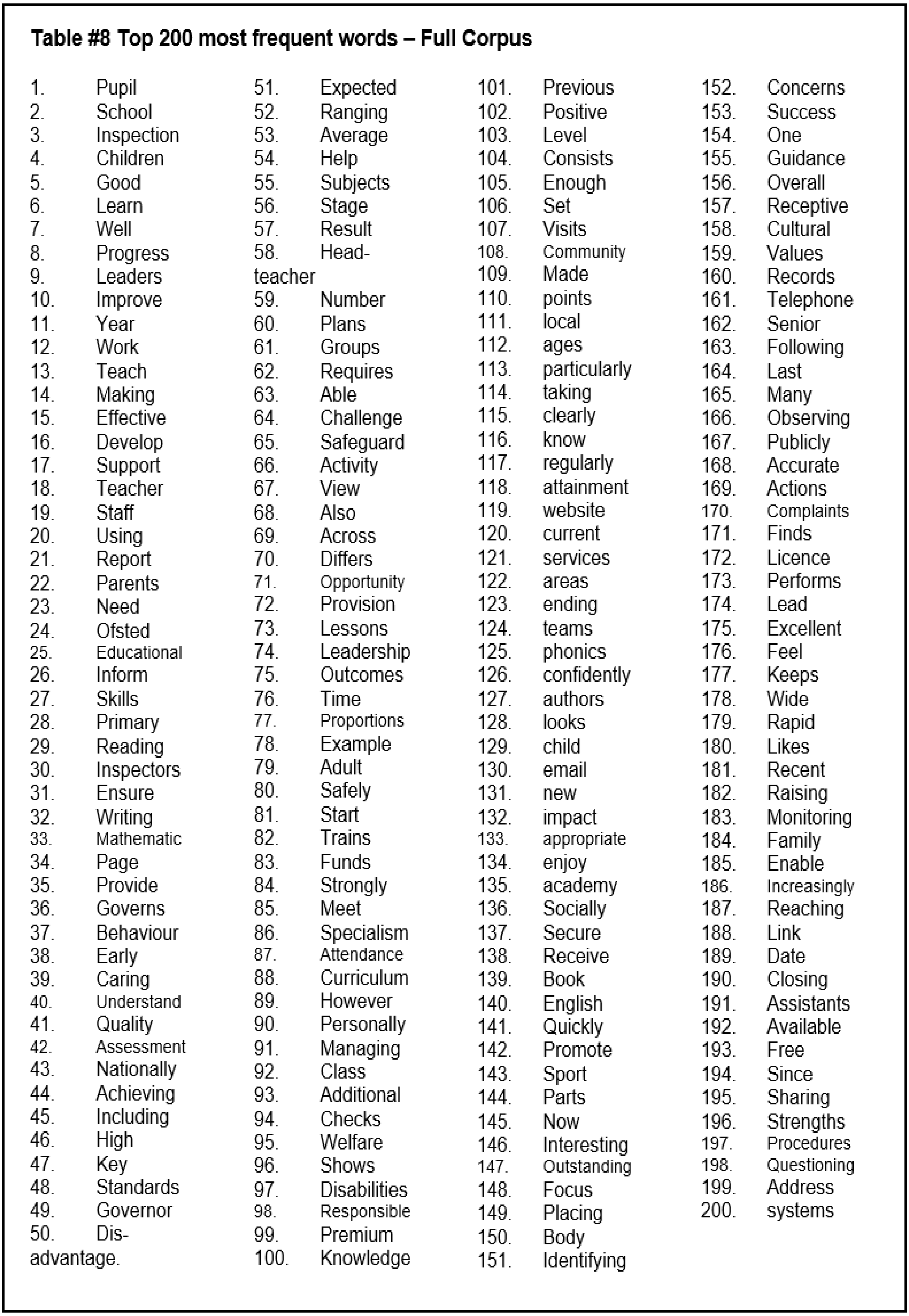


Table #8 Top 200 most frequent words – Full Corpus

1. Pupil
2. School
3. Inspection
4. Children
5. Good
6. Learn
7. Well
8. Progress
9. Leaders
10. Improve
11. Year
12. Work
13. Teach
14. Making
15. Effective
16. Develop
17. Support
18. Teacher
19. Staff
20. Using
21. Report
22. Parents
23. Need
24. Ofsted
25. Educational
26. Inform
27. Skills
28. Primary
29. Reading
30. Inspectors
31. Ensure
32. Writing
33. Mathematic
34. Page
35. Provide
36. Governs
37. Behaviour
38. Early
39. Caring
40. Understand
41. Quality
42. Assessment
43. Nationally
44. Achieving
45. Including
46. High
47. Key
48. Standards
49. Governor
50. Dis-advantage.
51. Expected
52. Ranging
53. Average
54. Help
55. Subjects
56. Stage
57. Result
58. Head-teacher
59. Number
60. Plans
61. Groups
62. Requires
63. Able
64. Challenge
65. Safeguard
66. Activity
67. View
68. Also
69. Across
70. Differs
71. Opportunity
72. Provision
73. Lessons
74. Leadership
75. Outcomes
76. Time
77. Proportions
78. Example
79. Adult
80. Safely
81. Start
82. Trains
83. Funds
84. Strongly
85. Meet
86. Specialism
87. Attendance
88. Curriculum
89. However
90. Personally
91. Managing
92. Class
93. Additional
94. Checks
95. Welfare
96. Shows
97. Disabilities
98. Responsible
99. Premium
100. Knowledge
101. Previous
102. Positive
103. Level
104. Consists
105. Enough
106. Set
107. Visits
108. Community
109. Made
110. points
111. local
112. ages
113. particularly
114. taking
115. clearly
116. know
117. regularly
118. attainment
119. website
120. current
121. services
122. areas
123. ending
124. teams
125. phonics
126. confidently
127. authors
128. looks
129. child
130. email
131. new
132. impact
133. appropriate
134. enjoy
135. academy
136. Socially
137. Secure
138. Receive
139. Book
140. English
141. Quickly
142. Promote
143. Sport
144. Parts
145. Now
146. Interesting
147. Outstanding
148. Focus
149. Placing
150. Body
151. Identifying
152. Concerns
153. Success
154. One
155. Guidance
156. Overall
157. Receptive
158. Cultural
159. Values
160. Records
161. Telephone
162. Senior
163. Following
164. Last
165. Many
166. Observing
167. Publicly
168. Accurate
169. Actions
170. Complaints
171. Finds
172. Licence
173. Performs
174. Lead
175. Excellent
176. Feel
177. Keeps
178. Wide
179. Rapid
180. Likes
181. Recent
182. Raising
183. Monitoring
184. Family
185. Enable
186. Increasingly
187. Reaching
188. Link
189. Date
190. Closing
191. Assistants
192. Available
193. Free
194. Since
195. Sharing
196. Strengths
197. Procedures
198. Questioning
199. Address
200. systems

Appendix 11 Words from the frequency list for full reports grouped into contextual areas using professional knowledge and insider insight.

Descriptors	Education Terms	Settings	Framework references
Good, well, Improve, Effective, Develop, Ensure Provide, Caring. Quality, Checks, Expected, Ranging, Average, Requires, Able, Challenge, Differs, Safely, Strongly, Meet, High Responsible, Previous, Positive, enough, particularly. clearly, regularly, current, ending, confidently, new, appropriate, enjoy, Socially, Secure, Quickly, Interesting, Outstanding, Wide, Success, Receptive, Cultural, Values, Last, Many, Publicly, Accurate, Performs, Excellent, Keeps, Rapid, Likes, Recent, Raising, Increasingly, Reaching, Closing, Strengths	Learn, Progress. Work, Teach. Making, Support Using, Educational Skills, Reading Writing, Maths Understand, Assessment Achieving Disadvantage Result, Number Plans, Trains Specialism Attendance Curriculum Knowledge Level, know. Phonics, Book English, Sport Complaints Monitoring Procedures Questioning, Sharing.	School Early Subjects Key Stage Provision Lessons Class Set Visits website services academy Records systems	Inspection Report Ofsted Inform Inspectors Nationally Standards Safeguard View Outcomes Proportions attainment impact Concerns Guidance Overall Outcomes Premium
Possible Stop Words	**Adults**	**Children**	**Others**
Year **Primary** **Page** **email** **Telephone** **Licence** **Date** **Address**	Leaders, Teacher Staff, Parents Governs, Governor. Headteacher Leadership Adult, Personally Managing, Community Local, teams Authors, Body Senior, Lead Family, Assistants	Pupil Children Behaviour Groups child	Need, Including. Help, Activity, Also, Across, Opportunity, Time, Example, Start, Funds, However, Additional, Welfare, Shows, Disabilities Consists, Made. Points, ages, Taking, Link, areas, Looks, Receive, Promote, Parts, Now, Focus, Placing, Identifying, One, Following, Observing, Actions, Finds, Feel, Enable Available, Free, Since

Appendix 12 Changes in word frequency ranking in full reports

Changes in word frequency ranking, from whole reports 'all' to 'deemed Section 5'.				
Small increase	Small decrease	Large increase	Large decrease	Gone from top 200
Assessment (+5) Achieve (+6) Stage (+4) Funds (+7) Premium (+4) Authors (+7) Overall (+8)	Good (-7) Behaviour (-3) Early (-5) Caring (-6) Adults (-4) Safely (-5) Attendance (-6) Enjoy (-6) Closing (-8)	Outstanding (+45) Excellent (+36) Recent (+24) Requires (+9) Enough (+13) Records (+11) Following (+9) Lead (+15) Systems (+11)	Now (-37) Previous (-23) Lessons (-11) Values (-10) Last (-9) Know (-11) Positively (-9)	Time (76th) Specialism (86th) Academy (135th) Assistant (192) Since (195)

Appendix 13 Changes in word frequency from just the 'area for improvement' section, all reports compared to 'deemed section 5' showing trends in terminology

Changes in word frequency ranking of 'areas for improvement' only from 'all' to 'deemed Section 5'.				
Small increase	Small decrease	Large increase	Large decrease	Gone from top 200
Effectiveness (+5) Managing (+7) Information (+9) Order (+6) Educational (+7) Meet (+7) Policy (+9) Receive (+9) Training (+10) Governors (+12) Governance (+12) Responsible (+12) Implement (+12) Outdoor (+12) Special (+12) Additional (+13) Clearly (+13)	Challenge (-4) Able (-5) Opportunity (-5) Highly (-5) Level (-6) Stage (-6) Handwriting (-6) Enable (-7) Particularly (-7) Phonics (-7) Spelling (-7) Better (-7) Reading (-8) Regularly (-8) Standards (-8) Curriculum (-8) Attainment (-8) Range (-8) Least (-9) Punctuation (-9) English (-10) Fully (-10) End (-11) Reduce (-11) Move (-11) Apply (-12)	Funding (+32) Hold (+31) Measurable (+28) Requirements (+27) Performing (+23) Undertaken (+22) May (+21) Targets (+20) Build (+20) Secure (+20) Systems (+19) Account (+18) Tracking (+18) Made (+18) Review (+17) External (+16) Premium (+15) Aspect (+15) Feedback (+15) Closely (+15) Classes (+14) Rigorously (+14)	Academy (-70) Deepen (-30) Depth (-28) Wider (-25) Reasoning (-23) Persistent (-23) Extend (-21) Attendance (-20) Lessons (-19) Continuing (-19) Especially (-16) Wide (-16) Higher (-16) Greater (-15) Strengthen (-15) Reach (-15) Grammar (-14)	Absence (193) Line (194) Age (198)

Appendix 14 A table showing the full T Test results for similarity % compared by did and did not improve progress and did and did not improve attainment.

	All maths areas for improvement	Did improve progress	Did not improve progress	Did improve attainment	Did not improve attainment
N	588	353	248	277	309
P		0.0075	0.379	0.0070	0.765
Mean	25.39%	27.82%	17.66%	28.73%	22.38%

Appendix 15 Anova results from single words and combinations

Anova: Single Factor Summary					
Groups	***Count***	***Sum***	***Average***	***Variance***	
All *(every report with maths results)*	586	425.7	1.040	10.315	
Includes problem	143	148.7	1.050	10.668	
Includes reasoning	133	139.6	2.023	7.010	
Includes accuracy	30	60.7	1.172	10.180	
Includes number	46	53.9	0.373	9.548	
Includes calculation	22	8.2	0.765	9.925	
Includes most able	375	286.8	0.831	10.490	
Includes challenge	194	161.3	1.184	9.265	
Includes expectations	204	241.5	0.726	9.524	
Does not include problem	443	277	0.625	9.249	
Does not include reasoning	453	286.1	0.632	9.171	
Does not include accuracy	556	365	0.656	9.576	
Does not include number	540	371.8	0.689	9.468	
Does not include calculation	564	417.5	0.740	9.535	
Does not include most able	211	138.9	0.658	8.847	
Does not include challenge	392	264.4	0.674	9.063	
Does not include expectations	382	184.2	0.482	9.515	
Includes expectations and challenge	75	64.8	0.864	11.798	
Includes expectations and most able	148	157.6	1.065	9.984	
Includes expectations and calculation	9	8.8	0.978	15.702	
Includes expectations and number	24	25.7	1.071	6.893	
Includes expectations and accuracy	16	40.1	2.506	6.763	
Includes expectations and reasoning	51	93.2	1.827	7.210	
Includes expectations and problem	62	108.8	1.755	8.112	
Includes challenge and most able	150	140.1	0.934	11.028	
Includes challenge and calculation	11	-1.7	-0.155	8.903	
Includes challenge and number	14	7.7	0.550	5.878	
Includes challenge and accuracy	9	25.7	2.856	5.758	
Includes challenge and reasoning	40	48.8	1.220	8.117	
Includes challenge and problem	46	38.7	0.841	8.946	
Includes most able and calculation	12	-2.3	-0.192	12.686	
Includes most able and number	28	30.5	1.089	13.387	
Includes most able and accuracy	20	42.4	2.120	8.615	
Includes most able and reasoning	85	99	1.165	12.465	
Includes most able and problem	101	114.6	1.135	11.920	
Includes calculation and number	8	2.9	0.363	3.994	
Includes calculation and accuracy	1	3	3	1	
Includes calculation and reasoning	5	-5.5	-1.100	10.220	
Includes calculation and problem	14	4.1	0.293	13.058	
Includes number and accuracy	3	-1.8	-0.600	1.750	
Includes number and reasoning	14	34.6	2.471	22.121	
Includes number and problem	14	16.4	1.171	21.356	
Includes accuracy and reasoning	10	23.3	2.330	11.833	
Includes accuracy and problem	11	15.7	1.427	8.692	
Includes reasoning and problem	92	86.6	0.941	11.569	
ANOVA Source of Variation	***SS***	***df***	***MS***	***F***	***P - value***
Between Groups	**605.6771**	**44**	**13.76539**	**1.420199**	**0.035312**
Within Groups	**61082.64**	**6302**	**9.69258**		
Total	**61688.32**	**6346**			

Appendix 16 T Test Results for progress – Full list of T tests where significance was found. Groups include where single words are present, combinations of words are present and groups of ‘not including’ are present and ‘all’ (full maths corpus) as a stand-alone group.

PROGRESS: where significance was found	**P**	**Mean group 1**	**Mean group 2**
Groups with single words			
accuracy / all	0.024	2.02	0.73
accuracy / calculation	0.044	2.02	0.37
accuracy / most able	0.034	2.02	0.76
accuracy / not accuracy	0.018	2.02	0.66
expectations v not expectations	0.009	1.18	0.48
Single words / group where two words are present (combination)			
accuracy / calculation and reasoning	0.023	2.02	-1.1
accuracy / challenge and calculation	0.030	2.02	-0.15
accuracy / most able and calculation	0.033	2.02	-0.19
all / challenge and accuracy	0.040	0.73	2.86
all / expectations and accuracy	0.023	0.73	2.5
all / expectations and problem	0.012	0.73	1.75
all / expectations and reasoning	0.014	0.73	1.83
all / most able and calculation	0.047	0.73	-0.19
all / most able and accuracy	0.047	0.73	2.12
all / number and reasoning	0.040	0.73	2.47
calculation / challenge and accuracy	0.040	0.37	2.86
calculation / expectations and accuracy	0.031	0.37	2.5
calculation / expectations and problem	0.059	0.37	1.75
calculation / expectations and reasoning	0.046	0.37	1.83
challenge / expectations and accuracy	0.045	0.83	2.5
challenge / expectations and problem	0.046	0.83	1.75
challenge / expectations and reasoning	0.044	0.83	1.83
expectations / expectations and accuracy	0.093	1.18	2.5
expectations / expectations and problem	0.190	1.18	1.75
expectations / expectations and reasoning	0.168	1.18	1.83
most able / challenge and accuracy	0.049	0.76	2.86
most able / expectations and accuracy	0.030	0.76	2.5
most able / expectations and problem	0.021	0.76	1.75
most able / expectations and reasoning	0.022	0.76	1.83
Combination of words group / combination of words group			
challenge and accuracy / calculation and reasoning	0.022	2.86	-1.1
challenge and accuracy / calculation and number	0.036	2.86	0.36
challenge and accuracy / most able and calculation	0.039	2.86	-0.19
challenge and accuracy / number and accuracy	0.042	2.86	-0.6
challenge and calculation / challenge and accuracy	0.025	-0.15	2.86
challenge and calculation / challenge and reasoning	0.025	-0.15	1.22
challenge and calculation / most able and accuracy	0.049	-0.15	2.12
challenge and number / challenge and accuracy	0.036	0.36	2.86

expectations and accuracy / calculation and reasoning	0.019	2.5	-1.1
expectations and accuracy / challenge and calculation	0.021	2.5	-0.15
expectations and accuracy / challenge and number	0.043	2.5	0.36
expectations and accuracy / most able and calculation	0.028	2.5	-0.19
expectations and problem / calculation and reasoning	0.036	1.75	-1.1
expectations and problem / challenge and calculation	0.046	1.75	-0.15
expectations and problem / most able and calculation	0.041	1.75	-0.19
expectations and reasoning / challenge and calculation	0.033	1.83	-0.15
expectations and reasoning / calculation and reasoning	0.026	1.83	-1.1
expectations and reasoning / most able and calculation	0.032	1.83	-0.19
most able and accuracy / calculation and reasoning	0.041	2.12	-1.1
Groups with ‘not included’	**P**	**Mean group 1**	**Mean group 2**
accuracy / not calculation	0.026	2.02	0.74
accuracy / not challenge	0.018	2.02	0.67
accuracy / not expectations	0.008	2.02	0.48
accuracy / not most able	0.018	2.02	0.66
accuracy / not number	0.020	2.02	0.69
accuracy / not problem	0.014	2.02	0.63
accuracy / not reasoning	0.014	2.02	0.63
expectations / not accuracy	0.037	1.18	0.66
expectations / not number	0.050	1.18	0.69
expectations / not problem	0.030	1.18	0.63
expectations / not reasoning	0.031	1.18	0.63
not accuracy / challenge and accuracy	0.034	0.66	2.86
not accuracy / expectations and accuracy	0.018	0.66	2.5
not accuracy / expectations and problem	0.008	0.66	1.75
not accuracy / expectations and reasoning	0.009	0.66	1.83
not accuracy / most able and accuracy	0.038	0.66	2.12
not accuracy / number and reasoning	0.033	0.66	2.47
not calculation / challenge and accuracy	0.041	0.74	2.86
not calculation / expectations and accuracy	0.024	0.74	2.5
not calculation / expectations and problem	0.014	0.74	1.75
not calculation / expectations and reasoning	0.015	0.74	1.83
not calculation / number and reasoning	0.042	0.74	2.47
not challenge / challenge and accuracy	0.032	0.67	2.86
not challenge / expectations and accuracy	0.017	0.67	2.5
not challenge / expectations and problem	0.008	0.67	1.75
not challenge / expectations and reasoning	0.010	0.67	1.83
not challenge / most able and accuracy	0.037	0.67	2.12
not challenge / number and reasoning	0.033	0.67	2.47
not expectations / challenge and accuracy	0.023	0.48	2.86
not expectations / expectations and accuracy	0.010	0.48	2.5
not expectations / expectations and problem	0.002	0.48	1.75
not expectations / expectations and reasoning	0.003	0.48	1.83
not expectations / most able and accuracy	0.021	0.48	2.12

not expectations / number and reasoning	0.021	0.48	2.47
not most able / challenge and accuracy	0.030	0.66	2.86
	P	**Mean group 1**	**Mean group 2**
not most able / expectations and accuracy	0.017	0.66	2.5
not most able / expectations and problem	0.011	0.66	1.75
not most able / expectations and reasoning	0.011	0.66	1.83
not most able / most able and accuracy	0.037	0.66	2.12
not most able / number and reasoning	0.035	0.66	2.47
not number / challenge and accuracy	0.036	0.69	2.86
not number / expectations and accuracy	0.020	0.69	2.5
not number / expectations and problem	0.009	0.69	1.75
not number / expectations and reasoning	0.011	0.69	1.83
not number / most able and accuracy	0.041	0.69	2.12
not number / number and reasoning	0.036	0.69	2.47
not problem / challenge and accuracy	0.029	0.63	2.86
not problem / expectations and accuracy	0.015	0.63	2.5
not problem / expectations and problem	0.006	0.63	1.75
not problem / expectations and reasoning	0.007	0.63	1.83
not problem / most able and accuracy	0.032	0.63	2.12
not problem / number and reasoning	0.029	0.63	2.47
not reasoning / challenge and accuracy	0.029	0.63	2.86
not reasoning / expectations and problem	0.006	0.63	1.75
not reasoning / expectations and reasoning	0.007	0.63	1.83
not reasoning / most able and accuracy	0.032	0.63	2.12
not reasoning / number and reasoning	0.029	0.63	2.47
not reasoning / expectations and accuracy	0.015	0.63	2.5

References

Adams, R (2016) Schools under scrutiny at Crackdown on league table 'gaming' *The Guardian* also Ozga, J., Dahler-Larsen, P., Segerholm, C. and Simola, H. (eds) (2011) *Fabricating Quality in Education: Data and Governance in Europe*, Abingdon: Routledge

Allen, G.C. (1960) 'H.M. Inspector of Schools: A personal impression', *International Review of Education*, 6(2): pp.235– 239.

Altrichter, H. & Kemethofer, D. (2015) Does accountability pressure through school inspections promote school improvement? *School Effectiveness and School Improvement,* 26:1. pp.32-56.

Andersen, V N, Dahler-Larsen, P. & Pedersen, C. S. (2009) Quality assurance and evaluation in Denmark, *Journal of Education Policy*, 24:2, pp.135-147

Archer-Kath, J., Johnson, D.W. and Johnson, R.T. (1994) Individual versus group feedback in cooperative groups, *Journal of Social Psychology*, 134 (5), pp.681– 694

Baker, K. (1993) *Those Turbulent Years: My Life in Politics*, London: Faber and Faber

Ball, S. (2003) The teacher's soul and the terrors of performativity, *Journal of Education Policy*, 18(2): pp.215– 228.

Ball, S.J. (1997) Good School/Bad School: paradox and fabrication, *British Journal of Sociology of Education,* 18:3, pp.317-336

Bazerman, C and Prior, P. (2004) *What writing does and how it does it: an introduction to analysing texts and textual practices.* Mahwah, N.J. Lawrence Erlbaum Associates.

Bazerman, Charles, and Paul A. Prior. *What Writing Does and How It Does It: an Introduction to Analysing Texts and Textual Practices*. Mahwah, N.J: Lawrence Erlbaum Associates, 2004.

BERA (2011) Ethical Guidelines for Educational Research. 1st Edition. London: British Educational Research Association.

Biesta, G (2013) *The beautiful risk of education* Boulder, CO. Paradigm Publishers

Biesta, G. (2009) Good Education in an Age of Measurement: on the need to reconnect with the question of purpose in education *Educational Assessment, Evaluation and Accountability* 21 (1) pp.33-46

Billig, M. (2002) *Critical discourse analysis and the rhetoric of critique*. In G. Weiss and R. Wodak (eds.), *Critical Discourse Analysis: Theory and Interdisciplinarity*. London: Palgrave Macmillan, pp.35-46.

Black, P. and William, D. (1998) Assessment and classroom learning, *Assessment in Education*, 5 (1), pp.7–75

Blacker, D.J. (2013) *The falling rate of Learning and the neoliberal endgame* Winchester, Zero Books and Pearce, J. (1986) School oversight in England and Wales, *European Journal of Education*, 21(4): pp.331– 344

Bloom, A (2017) Ofsted launches investigation into 'scandal' of schools gaming the system *Times Educational Supplement* https://www.tes.com/mews/school-news/breaking-news/fsted-launches-investigation-scandal-schools-gaming-system also in Ozga, J., Grek, S. and Lawn, M. (2009) *The new production of governing knowledge: Education research in the UK*, Soziale Welt, 60(4): pp.353– 370.

Braithwaite, J. (2008) *Regularity capitalism. How it works, ideas for making it work better.* Cheltenham: Edward Elgar.

Breeze, R. (2011). *Critical Discourse Analysis and Its Critics*. Pragmatics (214) p493-525

Brimblecombe, N., Shaw, M. and Ormston, M. (1996) Teachers' intention to change practice as a result of OFSTED School Inspections, *Educational Management & Administration*, 24 (4), pp.339–354

British Educational Research Association (BERA) (2018) Ethical guidelines for educational research. 4th edition.

Case, P. Case, S. & Catling, S. (2000) Please Show You're Working: A critical assessment of the impact of Ofsted inspection on primary teachers. *British Journal of Sociology of Education*, 21:4, 605-621, DOI: 10.1080/713655370

Chapman, C. (2001) Changing classrooms through inspection, *School Leadership and Management*, 21 (1), pp.59–73

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics*, 93, pp.1045–1057

Clarke, J. (2004) *Changing Welfare, Changing States*, Sage: London

Clarke, J. (2005) 'Producing transparency? Evaluation and the governance of public services', in Gavin Drewry, Carsten Greve and Thierry Tanquerel (eds) Contracts, *Performance Measurement and Accountability in the Public Sector*, International Institute of Administrative Science Monograph 25, Amsterdam: IOS Press, pp.41– 56.

Coffield, F (2009) *Ofsted Inspected: Adults Learning*, December 2009, Vol.21(4), p.26

Coffield, F and Williamson, B (2011) *From exam factories to communities of discovery; The Democratic Route.* In Bedford Way Papers, London, Institute of Education

Coffield, F. (2012) *Why the McKinsey reports will not improve school systems*, Journal of Education Policy, 27:1, 131-149, DOI: 10.1080/02680939.2011.623243

Courtney, (2013) Headteachers experience of inspection under the 2012 framework, *Management in Education* 27(4) p.164

Creasey, R. (2018) *The taming of education* Palgrave, Macmillan

Creemers, B. P. M., Stoll, L., & Reezigt, G. (2007). Effective school improvement – ingredients for success: The results of an international comparative study of best practice case studies. In T. Townsend (Ed.), International handbook of school effectiveness and improvement New York, NY: Springer. pp. 825–838.

Cresswell, J (2009 and 2014) Research design, qualitative, quantitative, and mixed method approaches (3rd ed) Thousand Oaks, CA: Sage.

Cresswell, J.W. and Plano Clark, V.L. (2011) *Designing and conducting mixed methods research* (2nd Ed) Thousand Oaks, CA: SAGE

Curtis, B (2012) *How the Ofsted Chief got his Maths wrong on SATS*. The Guardian https://www.theguardian.com/politics/reality-check-with-polly-curtis/2012/mar/15/ofsted-chief-maths-wrong Accessed July 2021

Dahler-Larsen, P. (2011) 'Afterword: Evaluation as a field and as a source of reflection: comments on how QAE restructures education now and in the future', in Ozga, J. Dahler-Larsen, P. Segerholm, C. and Simola, H. (eds) *Fabricating Quality in Education: Data and Governance in Europe*, Abingdon: Routledge, pp.150– 155.

Dastjerdi, A. V. (2016) *Big Data: Principles and Paradigms*, edited by Rodrigo N. Calheiros, and Rajkumar Buyya, Elsevier Science & Technology, 2016. ProQuest Ebook Central https://ebookcentral.proquest.com/lib/livhope/detail.action?docID=4544734

De Wolf, I. F., & Janssens, F. J. G. (2007). Effects and side effects of inspections and accountability in education: An overview of empirical studies. *Oxford Review of Education,* 33, pp.379–396

DFE (2013) Teacher standards URN https://www.gov.uk/government/collections/teachers-standards Accessed July 2021

DFE (2016) National curriculum assessments at key stage 2 in England. URN https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/549432/SFR39_2016_text.pdf Accessed July 2021

DFE (2016) Provisional KS2 publishing p31 *"We suppress some figures Values of 1 or 2, or a percentage based on 1 or 2 pupils who achieved; or 0, 1 or 2 pupils who did not achieve a particular level are suppressed. Some additional figures have been suppressed to prevent the possibility of a suppressed figure being revealed*" https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/577296/SFR62_2016_text.pdf Accessed March 2021

DFE (2016) SATS publications https://www.gov.uk/government/publications/key-stage-1-and-2-national-curriculum-tests-information-for-parents Accessed July 2021

DFE (2016) Understanding scaled scores URN https://www.gov.uk/guidance/understanding-scaled-scores-at-key-stage-2 Accessed March 2021

DFE (March 2019) Running Rural Primary Schools efficiently. URN https://www.gov.uk/government/publications/running-rural-primary-schools-efficiently accessed March 2022

DFE (March 2022) Opportunity for all: strong schools with great teachers for your child https://www.gov.uk/government/publications/opportunitiy-for-all-strong-schools-with-great-teachers-for-your-child Accessed April 2022

DFE scaled score guidance (2017) URN https://www.gov.uk/guidance/understanding-scaled-scores-at-key-stage-2 Accessed July 2021

DFE Website - schools financial benchmarking URN https://schools-financial-benchmarking.service.gov.uk Accessed July 2021

DFE Website - schools information URN https://get-information-schools.service.gov.uk Accessed July 2021

Dronkers and Veenstra, 2001 in *On the Impact of the Dutch Educational Supervision Act: Analysing assumptions concerning the inspection of primary education*, 2001 Ed. Ehren, M and Scheerens, J.

Dutch Chief Inspector Kervezee, quoted in Ehren and Visscher (2006) p51 Towards a theory on the impact of school inspections *British Journal of Educational Studies* 54:1 p54

Earley, P. (1998) S*chool Improvement after inspection? School and LEA responses*. London: Chapman, P. Pub.; Thousand Oaks, CA. Sage.

Education Act (2005) Available at https://www.legislation.gov.uk/ukpga/2005/18 Accessed September 2019

Education Act, Section 60 (4A) - (4B) of the Education and Inspections Act, as inserted by the Education and Adoption Act 2016, and Section 60(1)(d) and 60A(1)(d) of the Education and Inspections Act 2006 as amended by the Education and Adoption Act 2016.

Education and Inspection Act (*2006*) England, URN: http://www.legislation.gov.uk/ukpga/2006/40/ Accessed June 2021 Section 117 (1)(a) and Section 18

Ehren, M. C. M. and Visscher, A. J. (2006) Towards a theory on the impact of school inspections *British Journal of Educational Studies* 54:1 pp.54

Ehren, M. C. M., & Visscher, A. J. (2008). The relationships between school inspections, school characteristics and school improvement. *British Journal of Educational Studies*, 56, 2

Ehren, M. Perryman, J. & Shackleton, N. (2015) Setting expectations for good education: how Dutch school inspections drive improvement, *School Effectiveness and School Improvement*, 26:2, pp.296-327

Ehren, M.C.M, Gustafsson, J.E. Altrichter, H., Skedsmo, G., Kemethofer, D. & Huber, S.G. (2015) Comparing effects and side effects of different school inspection systems across Europe, *Comparative Education*, 51:3, pp.375-400

Ehrens (2014) *The impact of school inspections on teaching and learning (2014)* A Eurydice commissioned project. www.schoolsinspection.eu Project ISI-TL project number 511490-2010-LLP-NL-KA1-KA1SCR

Eurydice. (2004). Evaluation of schools providing compulsory education in Europe. URL http://www.eurydice.org/portal/page/portal/Eurydice Accessed September 2019

Exley, S. (Jan 9, 2015) *Ofsted and reliability* TES Issue 5128, London

Fairclough, N (1989) *Language and Power* London, Longman

Faubert, V. (2009). School evaluation: Current practices in OECD countries and a literature review *OECD Educational Working Papers, No. 42*. Retrieved from http://dx.doi.org/10.1787/218816547156

Finn, M. (2015) The politics of education revisited: Anthony Crosland and Michael Gove in Historical perspective *London Review of Education* v13 p98-112 Sept. and Education beyond the Gove legacy. In '*The Gove Legacy'* ed. Finn, M 14 Basingstoke; Palgrave, Macmillan

Fitzgibbon, C. (1999) *An Inspector Calls: Ofsted and its effect on school standards* ed. C Cullingford.

Fowler (1985) 'Power' in Van Dijk (ed) *Handbook of Discourse Analysis* vol 4 London, Academic Press p61

Fraenkel, J.R., Wallen, N.E. and Hyun, H. (2015) *How to design and evaluate research in education* New York: McGraw-Hill Education

Good, T., Wiley, C, Sabers, D (2019) Accountability and Educational Reform: A Critical Analysis of Four Perspectives and Considerations for Enhancing Reform Efforts, *Educational Psychologist*, v45 n2 pp.138-148

Government information on academies https://www.gov.uk/types-of-school/academies Accessed August 2021.

Government proposes "Ofsted style" ratings for CCG services *BMJ* 2015; 351 doi: https://doi-org.ezproxy.hope.ac.uk/10.1136/bmj.h5839 (Published 30 October 2015) *BMJ* 2015;351:h5839

Gray, C. and Gardner, J. (1999) The impact of school inspections, *Oxford Review of Education*, 25 (4), pp.455–469

Gray, J. and Wilcox, B. (1995) Good School, Bad School: Evaluating Performance and Encouraging Improvement Buckingham, Open University Press.

Greger, D. (2011). School inspection of Czech schools: A critical reflection on intended effects and causal mechanisms. Unpublished working paper, LLP project "Impact of school inspections on teaching and learning", Charles University, Prague.

Grek, S and Lindgren, J. (2014) Governing by Inspection. Routledge

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management* 24, pp.297–327.

Hartog, J. and Oosterbeek, H (2007) What you should know about the private return to education? In *Human Capital: advances in theory and evidence* ed. Hartog, J. and van den Brink, H. p7-20 Cambridge University Press

Hesse, B.W, Moser, R.P, and Riley, W.T (2015) *From Big Data to Knowledge in the Social Sciences* Annals of the American academy of political and social science Vol 659 (May 2015) pp.16-32

Hirsch, E.D Jr (1967) *Validity in interpretation* New Haven / London Yale University Press

HMCI report (2015), England. URN: https://www.gov.uk/government/collections/ofsted-annual-reports Accessed September 2019

House of Commons Hansard Debates for 4 Jul 2002 (part 1) (parliament.uk) and Ofsted and Spielman, A. (2017) Speech at the ASCL Annual Conference. HM Govt. [cited 24 March 2017] https://www.gov.uk/government/publications/school-inspection-handbook-eif Accessed July 2021 https://www.gov.uk/government/speeches/amanda-spielmans-speech-at-the-ascl-annual-conference

Husfeldt, V. (2011). The impact of school inspection – does it really work? State of research. *Zeitschrift für Erziehungswissenschaft*, 14, pp.259–282

Hussey, T. and Smith, P. (2010) *The trouble with Higher Education* Abingdon, Routledge

Hyatt, D. (2005b). Time for a change: A critical discoursal analysis of synchronic context with diachronic relevance. *Discourse in Society, 16*(4), pp.515-534.

Hyatt, D. (2014). The critical higher education policy discourse analysis framework. *Theory and method in higher education research*, *9*, pp.41-59.

Hyatt, D., & Meraud, J. (2015). Teacher education in France under the Hollande government: Reconstructing and reinforcing the republic. *Journal of Education for Teaching: International research and pedagogy,* 1-17. http://dx.doi.org/10.1080/02607476.2015.1044227

Iacobucci, G. "*GP inspections: are sanctions holding back improvement in poor areas?*" *BMJ* 2018; 360 (Published 14 February 2018) *BMJ* 2018; 360: k682

Ilgen, D.R., Fisher, C.D. and Taylor, M.S. (1979) Consequences of individual feedback on behaviour in organizations, *Journal of Applied Psychology*, 64 (4), pp.349–371.

Janssens, F and Maassen, N. (2014) School Inspections in a polycentric context: The Dutch Inspectorate of Education. From *The impact of school Inspection on teaching and learning* Erasmus 2011-2014

Janssens, F and van Amelsvoort, G. (2008) School self-evaluations and school inspections in Europe: An exploratory study *Studies in Educational Evaluation* Vol 34 (2008) pp.15-23

Jaworski, A, and Coupland, N. (eds) (1999) *The Discourse Reader* New York, Routledge

Jones, K. & Tymms, P. (2014) Ofsted's role in promoting school improvement: the mechanisms of the school inspection system in England, *Oxford Review of Education*, 40:3, 315-330, DOI: 10.1080/03054985.2014.911726

Karsten, S. and Visscher, A.J. (2001) Publishing school performance indicators: some lessons, *Journal of behavioural, management and social sciences*.

Kesby, M. (2005) Retheorizing Empowerment-through-Participation as a performance in space: beyond tyranny to transformation

Klerks, M. (2013). The effect of school inspections: A systematic review. Manuscript submitted for publication.

Kogan, M. (1971) *The Politics of Education*, London: Macmillan Education.

Kogan, M. and Maden, M. (1999) An evaluation of evaluators: the OFSTED system of school inspection. In C. CULLINGFORD (Ed.) *An Inspector Calls; Ofsted and its Effect on School Standards* London, Kogan Page, pp.9–32

Kotthoff, H.-G., & Böttcher, W. (2010). New forms of "school inspection ". Empirical education research on expectations of effects and effectiveness. In H. Altrichter & K. Maag Merki (Eds.), Handbuch Neue Steuerung im Schulsystem (pp. 295– 325). Wiesbaden: VS Verlag für Sozialwissenschaften.

Lawn, M. (2006) Soft governance and the learning spaces of Europe, *Comparative European Politics*, 4(2): 272– 288 and Lawn, M. and Grek, S. (2012) *Europeanizing Education: Governing a New Policy Space*, Oxford: Symposium Books

Lazar, M.M. (2005) *Feminist Critical Discourse Analysis: Gender, Power and Ideology in discourse.* Houndmills, UK Palgrave Macmillan.

Lee, J. and Fitz, J. (1998) HMI and OFSTED: Evolution or revolution in school inspection, *British Journal of Educational Studies*, 45(1): pp.39– 52.

Leeuw, F.L. (1995) External independence and accountability information, op 9 June 1995, *Openbare uitgaven*, 27 (4), 185–190. and Leeuw, F.L. (2000) Unexpected side-effects of output-steering, control, and supervision, Bijlage 6 uit het advies 'aansprekend burgerschap' van de Raad voor maatschappelijke ontwikkeling (The Hague, SDU Uitgevers).

Lefstein (2013) The regulation of teaching as symbolic politics rituals of order, blame and redemption *Discourse: Studies in the Cultural Politics of Education* 2013 34 (5) p643-659

Leite, et al (2014) Curriculum contextualisation: A comparative analysis of meanings expressed in Portuguese and English school evaluations. *Studies in Educational Evaluation* Vol 43, Dec 2014 pp.133-138

Lester, J. N., Lochmiller, C. R., & Gabriel, R. (2016). *Locating and applying critical discourse analysis within education policy: An introduction*. Education Policy Analysis Archives, 24 (102). http://dx.doi.org/10.14507/epaa.24.2768

Lindgren, J. Hult, A. Segerholm, C. & Rönnberg, L. (2012) Mediating school inspection, *Education Inquiry*, 3:4, 569-590, DOI: 10.3402/edui.v3i4.22055

London, M. (1995) Giving feedback: source-centered antecedents and consequences of constructive and destructive feedback, *Human Resource Management Review*, 5 (3), pp.159-188.

Luginbuhl, R., Webbink, D., & De Wolf, I. (2009). Do school inspections improve primary school performance? *Educational Evaluation and Policy Analysis*, 31, pp.221–237.

Maag Merki, K. (2010). *Theoretical and empirical analyses of the efficiency of educational standards, standard-related tests and central final examinations*. In H. Altrichter & K. Maag Merki (Eds.), Handbuch Neue Steuerung im Schulsystem (pp. 145–169). Wiesbaden: VS Verlag für Sozialwissenschaften.

Maclure, 1998 "What is the role of the inspectorate in respect of national standards and their maintenance" in Maclure, S. (1998) Through the revolution and out the other side, Oxford Review of Education, 24(1): pp.5– 24.

Martin, J. (2004) Positive discourse analysis: Solidarity and change. Revista Canaria de Estudios Ingleses 49: pp.179-202.

Matthews, P. and Sammons, P. (2004) *Improvement through Inspection* London, Ofsted

McPherson, A. and Raab, C.D. (1988) *Governing Education: A Policy Sociology of Education in Scotland*, Edinburgh: Edinburgh University Press

Means, Alexander J. Education for a post-work future: automation, precarity, and stagnation. *Knowledge Cultures*, vol. 5, no. 1, Jan. 2017, p. 21. *Gale Academic OneFile*

Mogra, I. The 'Trojan Horse' Affair and Radicalisation: An Analysis of Ofsted Reports. *Educational review (Birmingham)* 68.4 (2016): pp.444–465.

Morris, C. (1971). Writings on the General Theory of Signs. The Hague, Paris: Mouton. [Part 1: *Foundations of the Theorie of Signs* (1938):17-72; Part 2: *Signs, Language, and behaviour* (1946) pp.75-368

National Audit office (2018) URN https://www.nao.org.uk/report/ofsteds-inspection-of-schools/ Accessed July 2021

Nunez-Perucha, B. (2014) Critical Discourse Analysis and Cognitive Linguistics as tools for ideological research: a diachronic analysis of feminism in Critical discourse studies in *Context and Cognition*, Hart. C (ed) 2014

NVIVO manual re: autocode (2020) URN https://help-nv.qsrinternational.com/20/win/Content/coding/automatic-coding-existing-patterns.htm?Highlight=autocode Accessed July 2021

NVIVO manual re: stop words (2020) URN https://help-nv.qsrinternational.com/20/win/Content/queries/text-content-language-and-stop-words.htm Accessed July 2021

Ofsted https://parentview.ofsted.gov.uk Accessed July 2021

Ofsted https://reports.ofsted.gov.uk Accessed July 2021

Ofsted (2013) *"The Framework for School Inspection"* URN: https://www.gov.uk/government/publications/school-inspection-handbook-eif Accessed July 2021

Ofsted (2014) "*Reports must be written in clear, simple language so that the lay reader or parent can understand them. They must be jargon free. The main findings, strengths, weaknesses, and recommendations should be clearly spelled out so that there is no doubt about what the school needs to do to improve, or to maintain already outstanding practice. Avoid the unnecessary regurgitation of the language of the inspection handbook or guidance when writing reports. It adds little and often leads to bland reporting*". Introduction, bullet 2 p4 of Guidance for inspectors writing a section 5 inspection report published 4 April 2014, No. 120204

Ofsted (2017) Complaints Guidance https://www.gov.uk/complain-ofsted-report

Ofsted (2017) School Inspection Handbook #31. *"…Inspectors must not advocate a particular method of planning, teaching or assessment. It is up to schools themselves to determine their practices and for leadership teams to justify these on their own merits rather than by reference to this inspection handbook*." p12

Ofsted (2017) School Inspection handbook p10 *"The frequency of inspection is proportionate to the performance and circumstances of schools. Schools judged to be good at their previous section 5 inspection will normally receive a one-day short inspection, carried out under section 8, approximately every four years, as long as the quality of education remains good at each short inspection".* Available at URN https://www.gov.uk/government/publications/school-inspection-handbook-from-september-2015 Accessed July 2021

Ofsted (2017) The School Inspection handbook p47 *"The broad and balanced curriculum inspires pupils to learn".* Available at URN https://www.gov.uk/government/publications/school-inspection-handbook-from-september-2015 Accessed July 2021

Ofsted (2017) Inspection Handbook. Available at URN https://www.gov.uk/government/publications/school-inspection-handbook-from-september-2015 Accessed July 2021

Ofsted (2017) Outcomes report for 2017 https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/692223/Maintained_schools_and_academies_inspections_and_outcomes_as_at_31_December_2017.pdf Accessed July 2021

Ofsted (2017) School Inspection Handbook Available at URN https://www.gov.uk/government/publications/school-inspection-handbook-from-september-2015 Accessed July 2021

Ofsted (2017) Section 8 handbook https://www.gov.uk/government/publications/section-8-school-inspection-handbook-eif Accessed July 2021

Ofsted (2018) State funded school inspections and outcomes as at 31 March 2018 https://www.gov.uk/government/statistics/state-funded-schools-inspections-and-outcomes-as-at-31-march-2018/state-funded-schools-inspections-and-outcomes-as-at-31-march-2018-main-findings Accessed September 2019

Ofsted (2018) Diagram taken from URN https://www.gov.uk/government/organisations/ofsted/about/statistics

Ofsted (2018) Official Statistics: *Maintained schools and academies inspections and outcomes as at 31 December 2017*, published 22 March 2018.

Ofsted (2018) School inspection handbook No.150066

Ofsted (2018) School inspection handbook No.150066 #150 "*knowledge of Britain's democratic parliamentary system and its central role in shaping our history and values, and in continuing to develop Britain… Willingness to participate in and respond positively to artistic, musical, sporting and cultural opportunities*"

Ofsted (2018) School inspection handbook No.150066 *#60 "An increasing number of schools are adopting mastery approaches to the teaching of mathematics. Such approaches reflect particular beliefs and pedagogical practices. However, it is for each school to determine, in the best interests of its pupils, how the mathematics curriculum is taught"*.

Ofsted (2018) Statistical release March 2018. URN: https://www.gov.uk/government/collections/maintained-schools-and-academies-inspections-and-outcomes-official-statistics. Accessed September 2019

Ofsted (2019) New framework handbook

Ofsted report (2017) #10019995

Ofsted report (2017) #10033969

Ofsted report (2017) #10019298

Ofsted report (2017) #10024104

Ofsted report (2017) #10025177

Ofsted report (2017) #10025640

Ofsted report (2017) #10031364

Ofsted report (2017) #10031717

Ofsted report (2017) #10031772

Ofsted report (2017) #10032972

Ofsted report (2017) #10036618

Ofsted report (2017) #10036781

Ofsted report (2017) #10037735

Ofsted report (2017) #10042651

Ofsted report (2017) #10042802

Ofsted report (2017) #118025

Ofsted, (2019 update) https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/814685/Inspecting_the_curriculum.pdf Accessed June 2021

Ofsted, (2019) The nuance of reliability studies https://educationinspection.blog.gov.uk/2019/08/20 (accessed November 2021)

Ofsted, (2020) Annual Report of Her Majesty's Chief Inspector of education, Children's services, and Skills 2019/2020 U.K Parliament.

Ofsted, 2012 Subsidiary guidance April 2012 #110166 Describing 'Most' (80-95%), 'substantial proportion' (more than 20%) etc. Part 1, P6, bullet #8 URN https://www.gov.uk/government/collections/ofsted-inspections-of-maintained-schools Accessed July 2021

Ouston et al (1997) What do schools do after Ofsted or before? *School Leadership and Management* 17 (1) pp.95-104

Ouston, J. Fidler, B. and Earley, P. (1997) What do schools do after Ofsted or before? *School Leadership and Management* 17 (1) pp.95-104

Ozga, J. and Lawn, M. (2014) Inspectorates and Politics: the trajectories of school inspection in England and Scotland. *Revue française de pédagogie* no 186 pp.11-21.

Parr, A (2020) Children and Teachers All Felt He Was a Friend: the early years of Her Majesty's Inspectorate of Schools, 1837-70, FORUM, 62(3), 469-476. https://doi-org.ezproxy.hope.ac.uk/10.15730/forum.2020.62.3.469

Petty, N. J., Thomson, O. and Stew, G. (2012) Ready for a paradigm shift? Part 1: Introducing the philosophy of qualitative research Manual Therapy 17 (2012) p267-274

Phillips, J (1999) Not Learning but Coasting. *Checklist for school governors in Great Britain* 4309: 30

Pilkington, A (1991) Poetic effects: a relevance theory perspective In Sell, R.D (Ed) *Literary Pragmatics.* London, New York: Routledge pp.44-61

Power, M (2003) Evaluating the adult explosion *Law and Policy 25* (3) pp.185 - 202

Richards, C (2012) Ofsted Inspection Inspected: an examination of the 2012 framework and its accompanying evaluation schedule *Forum* Vol 54 No 2 2012, and Richards, C. (2016) A Tale of Two Interpretations: Ofsted's expectations re-examined *Forum* Vol 58, No 2, 2016

Rosenthal, L. (2004) Do school inspections improve school quality? Ofsted inspections and school examination results in the UK. *Economics of Education Review*, 23 (1) pp.143–151

Sakr, S (2016). *Big Data 2.0 Processing Systems A Survey*. Cham: Springer International Publishing, Web.

Scheipl, J., & Seel, H. (1985). Die Entwicklung des österreichischen Schulwesens von 1750 bis 1938 [The development of the Austrian education system – 1750 to 1938]. Graz: Leykam

Scholtes, E., Waslander, S. and Zoontjens, P. (2002) Is stimulating inspection more than government advice? In Eijander, P., Geste, Van R., Ligthart, W. and Waslander, S. Dilemmas of inspections: writing about inspection arrangements, The Hague, SDU Uitgevers.

Schweinberger, K. Quesel, C. Mahler, S. and Hochli, A. (2017) Effects of feedback on process features of school quality: A longitudinal study on the effects of teachers' reception of school inspection of Swiss compulsory schools. *Studies in Educational Evaluation* 55 (2017) pp.75-82.

Segerholm, C and Åström, E (2007) Governance through institutionalised evaluation: Re-centralisation and influences at local levels in higher education in Sweden, *Evaluation*, 13(1): pp.48– 67

Shaw, I., Newton, D.P., Aitkin, M. and Darnell, R. (2003) Do OFSTED inspections of secondary education make a difference to GCSE results? *British Educational Research Journal*, 29 (1)

Sin, K. and Muthu, L (2015) *application of big data in education data mining and learning analytics - a literature review* ICTACT Journal on soft computing (Jul 2015) Vol 5. No 4. pp.1035-1049.

Stobart, G (2008) *Testing Times: the uses and abuses of assessment* Abingdon: Routledge

Swedish inspectorate (2010) URN: https://files.eric.ed.gov/fulltext/EJ1075811.pdf Accessed July 2021

Syed, M. (2020) Rebel Ideas, the power of diverse thinking, John Murray ISBN-10: 1473613949, and Van Dijk (2009) *Discourse Studies* (2nd Edition) SAGE, London.

Thrift, N (2005) *Knowing Capitalism*, London Sage

Times Educational Supplement November 2017 Pupils being harmed by schools 'gaming' the system to climb league tables. Available at www.tes.com/news/pupils-being-harmed-schools-gaming-system-climb-league-tables accessed September 2019

Trena, P. et al (2017) *The discourse of QDAS: reporting practices of ATLAS.ti and NVivo users with implications for best practices*. International journal of Social Research Methodology, 02 Jan 2017 Vol 20(1) pp.35-47

Van Bruggen, J. C. (2010) Inspectorates of education in Europe; some comparative remarks about their tasks and work (SICI report). URL: http://www.sici-inspectorates.eu/getattachment/c2bfe3ff-49b7-4397-ae65-d0a203451928 Accessed July 2021

Van Dijk (2009) *Discourse Studies* (2nd Edition) SAGE, London, *Analysing Discourse: Textual Analysis for Social Research* London: Routledge

Van Dijk, T. A. (1993). Principles of critical discourse analysis. *Discourse & Society*, 4, pp.249-283

Van Dijk, T. A. (2001). Critical discourse analysis. *The handbook of discourse analysis*, pp.349-371

Van Dijk, T.A. (2008) *Discourse and Power,* Palgrave; 2008 edition

Waldegrave, H. and Simons, J. (2014) 'Watching the watchmen' Policy Exchange report *the future of school inspection in England*

Wang, H. (2015) A corpus-based contrastive study of online news reports on economic crisis - A critical discourse analysis perspective. ISSN 1798-4769 *Journal of Language Teaching and Research* Vol. 6 No. 3 pp 627-632, May 2015 DOI: http://dx.doi.org/10.17507/jltr0603.20

Wang, Y. (2017). Education policy research in the big data era: Methodological frontiers, misconceptions, and challenges. Education Policy Analysis Archives, 25(94). http://dx.doi.org/10.14507/epaa.25.3037

Watson, J. (2001) OFSTED's Spiritual Dimension: an analytical audit of inspection reports *Cambridge Journal of Education* Vol 31 (2) 2001

Weaver, (1979) *British Journal of Educational Studies*, 46(4): pp.415– 427 and Weaver, T. (1979) Department of Education and Science: Central control of education? Unit 2 of Course E222 *The Control of Education in Britain*, The Open University, Milton Keynes

Wiebes, F. (1998) From outside to inside; valuable ideas of outsiders, The Hague, Delwel Uitgeverij B.V.

Wiggins, K (2015) *Good' schools may be 'coasting' too.* Times Educational Supplement. Times Supplements Ltd. London

Wilcox, B. and Gray, J. (1996). *Inspecting Schools: Holding Schools to Account and Helping Schools to Improve* Buckingham / Philadelphia, University Press

Wimsatt, W.K and Beardsley, M.C. (1946) The intentional fallacy *the Sewanee Review* 54: pp.468-488

Wood, E. (2019) Unbalanced and unbalancing acts in the Early Years Foundation Stage: a critical discourse analysis of policy-led evidence on teaching and play from the Office for Standards in Education in England (Ofsted*). Education* 3-13, 47:7 pp.784-795

Woolf, N.H. and Silver, C. (2017) *Qualitative analysis using NVivo: the Five-Level QDA Method*. Routledge, 2017.

Word readability function. URN https://www.bettercloud.com/monitor/the-academy/find-readability-score-word-document/ Accessed July 2021

www.gov.uk/government/collections/ofsted's-plans-2021 (accessed November 2021)

Ofsted Inspection framework (2021) Education inspection framework - GOV.UK (www.gov.uk)