The unintended consequences of school inspection: The prevalence of inspection side-effects in Austria, the Czech Republic, England, Ireland, the Netherlands, Sweden and Switzerland.

Authors
Karen L. Jones, Durham University, England
Peter Tymms, Durham University, England
David Kemethofer, Johannes Kepler University Linz, Austria
Joe O'Hara, Dublin City University, Ireland
Gerry McNamara, Dublin City University, Ireland
Stephan Huber, University of Teacher Education Zug, Switzerland
Eva Myrberg, University of Gothenburg, Sweden
Guri Skedsmo, University of Teacher Education Zug, Switzerland
David Greger, Charles University in Prague, Czech Republic

Corresponding Author
Karen L. Jones
Postal Address: CEM, Durham University, Rowan, Stockton Road, Durham, UK, DH1 3UZ
Email: Karen.jones@cem.dur.ac.uk

Acknowledgements
This work was supported by the European Commission, Lifelong Learning Programme, under Grant DG EAC/41/09.

Abstract
It has been widely documented that accountability systems, including school inspections, bring with them unintended side-effects. These unintended effects are often negative and have the potential to undo the intended positive effects. However the empirical evidence is limited. Through a European comparative study we have had the rare opportunity to collect empirical evidence and study the effects (both intended and unintended) of school inspections (a key system of accountability) in a systematic way, across seven countries. We present the findings of the unintended effects in this paper. Survey self-report responses from school principals in each country, with differing school inspection systems, are analysed to measure the prevalence of these unintended effects and to investigate the part played by pressure to do well in inspections. A key finding is that increasing pressure in school inspection systems is associated with the undesired effect of the narrowing and refocussing of the curriculum and instructional strategies. We also show that a proportion of school principals admit to misrepresenting the school in data sent to the inspectorate and show evidence for formalisation/proceduralisation (excessive focus on records) and ossification (fear of experimentation in teaching), although these factors are less related to changes in pressure.

Key words: inspection, accountability, unintended effects, comparative research

Introduction

Systems of accountability, aimed at ensuring the quality of education in schools, now exist in the majority of countries across Europe (Eurydice 2004) and are widespread globally. Most of these systems take the form of school inspectorates, but they also involve the use of examination results and league tables. Since the work of Smith (1995) and Fitz-Gibbon (1997) in the 1990s, it is now widely accepted that accountability systems bring with them unintended effects, such as gaming, tunnel vision and measure fixation. These unintended effects are often negative, and as Leeuw (2000) suggests, may be so problematic as to undo the intended positive effects. However the empirical evidence for the unintended effects of school inspections and accountability systems is limited and De Wolf and Janssens (2007) conclude in their overview of the literature that more quality empirical research is needed. We aim to help fill this gap by studying the unintended consequences of school inspection systems in a systematic way, by analysing various consequences across a number of different countries. We present self-report survey data from school principals in seven countries with differing inspection systems in order to examine the apparent prevalence of a number of unintended effects such as narrowing of the curriculum. We show how these effects vary across the countries and suggest that these differences are related to the degree of pressure associated with the various inspection regimes. We argue that the prevalence and intensity of the unintended consequences of school inspections increases as the pressure exerted by the inspection system increases. We note that inspection systems are sometimes inextricably linked to examination results and league tables (and where this is the case we have considered the unintended consequences of national testing and league tables as an integral part of the inspection system).

The empirical data presented in this paper were collected as part of the European Commission's, Lifelong Learning Project, 'The impact of school inspection on teaching and learning' (ISI-TL). The project as a whole aims to fill gaps in knowledge about the impact of school inspection by comparing inspectorates in seven countries (England, the Netherlands, Ireland, Czech Republic, Austria, Sweden and Switzerland as an associate project partner). The study aims to identify aspects of school inspections (e.g. standards and thresholds, sanctions and rewards) that maximise the positive, intended effects of school inspection and minimise the negative, unintended effects (Ehren et al. 2013). This paper focuses on the survey items aimed at measuring the side effects of school inspection.

Due to the varying range of inspection systems, as part of the project, conceptual models were produced in each country to describe the inspectorate (see for example Jones and Tymms 2014) and were summarised in Ehren et al. (2013). Altrichter and Kemethofer (2015) used a priori data from these conceptual models to categorise each country's inspectorate according to how much pressure the schools and principals feel to do well in the inspection. England and the Netherlands were found to be high pressure inspection systems, Sweden, Ireland and the Czech Republic had medium pressure systems and Switzerland and Austria had the lowest pressure inspection systems. Evidence for this theoretical ordering was triangulated using a survey item designed to capture how much pressure principals feel under to do well on the inspection standards. In this paper we use these data to investigate unintended effects of school inspections.

The following research questions are addressed:

1. *What is the prevalence of unintended consequences of school inspections across seven European countries?*

2. *What part does pressure play in precipitating these unintended consequences?*

This paper begins by presenting a theoretical framework concerning the relationship between pressure and accountability systems. We then summarise the literature on unintended consequences of school inspection and go on to briefly describe the accountability systems in each of the countries in our project. This is followed by a description of the method and survey instrument used and details of the samples. We then present limitations, threats and mitigation followed by the results and end with a discussion and summary of the results.

Theoretical framework: schools responding to pressure

It is well known that stress can be employed to improve the performance of systems and that there can be unintended consequences. The study of the consequences of stress has been extensive in the physical sciences where perhaps the most relevant finding, originating from the nineteenth century, is Le Châtelier's principle (Thomsen 2000). It states that *"if a stress is applied to a system in equilibrium it will move in such a way as to negate the effect of stress".*

The principle is widely used in chemistry to predict the consequences of changing such variables as temperature and concentration but it might equally be extrapolated to the social sciences when systems are put under stress. The key point is that the system will seek to reduce the effect of stress; exactly how may not be as clear in the social world as in the test tube but one can predict that the system will change.

Within the social world it is widely accepted that as rules and regulations are put in place to improve performance, things often happen which were not intended. This was termed 'unintended consequences' by Merton (1936) and his heavily cited article suggested that these unintended consequences can be positive or negative. Within the social sciences Campbell's Law, paralleling Goodhart's Law (1975) from economics, has become well cited with reference to the unintended effects of educational accountability:

> *The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor.*
>
> Campbell (1976, 49)

These negative unintended consequences have been documented by Smith (1995) who outlined a series of ways in which groups within a system might behave in order to improve their indicators when put under stress. None of the actions were intended to improve performance but rather to give the appearance of improved performance. Fitz-Gibbon (1997) outlined how Smith's ideas could be applied to schools when put under pressure. De Wolf and Janssens (2007) extend these ideas and summarise the possible undesirable effects into three categories. Firstly, there is 'intended strategic behaviour and gaming'. This includes 'window dressing' where arrangements are made simply to make the school look more effective to the inspectorate. It also includes 'reshaping the test pool' and deliberate 'misinterpretation', for example with selective data reporting. At its extremes this category includes 'fraud and direct cheating', for example helping students in examinations or changing test scores. The second category is 'unintended strategic behaviour' where the behaviour of those being assessed is unintentionally changed by the process of inspection. Such side effects include 'formalisation and proceduralisation' (over-focus on records), concentrating on assessed elements 'and teaching to the test/inspection' leading to a narrowing of the curriculum and ossification (a fear of experimenting with new teaching methods). Such 'tunnel vision and indicator fixation' can lead to 'myopia' where schools focus on short-term solutions, rather than long term educational gains. An example of the impact of this category is offering easier exam syllabuses or subjects. De Wolf and Janssens (*ibid*) suggest that these unintended effects may lead to 'convergence and isomorphism', where all school become similar. Finally, the third category is 'other side effects' such as 'stress' felt by teachers and principals and good schools 'resting on their laurels' (Perryman 2007).

The research outlined above provides a theoretical framework which suggests that as pressure is gradually applied to school systems through inspection, the systems will respond in ways that were unforeseen and unintended and that some of these consequences are likely to be negative and damaging. This impact will vary across systems depending on the type and intensity of the pressure applied but it is theorised that as pressure rises, unintended negative consequences will start to appear and will increase in number and intensity.

Empirical evidence on the unintended effects of school inspection and accountability

De Wolf and Janssens (2007, 382) state that 'it is widely accepted that control systems have side effects'. However remarkably few empirical studies exist regarding the prevalence of these side effects in education, and there are large gaps in the literature, both in terms of the effects that have been studied and the countries in which they have been studied.

The majority of the studies in this area focus on intended strategic behaviour, investigating the prevalence of window dressing, misrepresentation, fraud and deception and reshaping the test pool. The evidence for window dressing primarily comes from studies in England, where Brimblecombe et al. 1996; Wilcox and Gray 1996; Fitz-Gibbon and Stephenson-Forster 1999; Case et al. 2000 and Chapman 2001 all find evidence that a proportion of schools 'put on a show' for inspectors. Fitz-Gibbon and Stephenson-Forster (1999) found that 81% of principals surveyed claimed that inspectors did not see the school in its normal state and Brimblecombe et al. (1996) found from their survey of teachers that a third suggested that the inspector did not see a typical lesson. Ehren (2006) found evidence of misrepresentation in Dutch schools where schools included outside playtime in the lesson schedules in order to comply with the minimum number of lesson hours. Evidence for fraud and cheating was also found. Ehren and Swanborn (2012) found that despite the fact that teachers were being observed by inspectors, some form of cheating behaviour/non-compliance with the rules was found in more than 5% of schools, ranging from clarifying test questions to prompting students with the correct answer. Jacob and Levitt (2003) investigated the prevalence of cheating by teachers in Chicago schools by analysing unusual score fluctuations and patterns of answers from students. They detected cheating by altering test paper answers in approximately 4-5% of classes, which due to the clandestine nature of the acts is likely to be an underestimate.

Many states in the US introduced accountability systems based on student achievement in the 1990s culminating in the 'No Child Left Behind' Act introduced in 2001. The Act requires states to judge the performance of schools based on annual test scores, which has led to a number of studies in the US focussing on the effects and side effects of performance indicators. Most of the studies on side-effects focus on measuring the extent to which schools are reshaping their test pool in order to improve their test scores. Jacob (2005), Cullen and Reback (2006) and Figlio and Getzler (2006), in Chicago, Texas and Florida respectively, all find evidence that since the publication of test scores there has been an increase in the reclassification of poorly performing students as having special educational needs. Many of these students are then exempt from the tests or from the test results being used to evaluate school performance. In contrast, Hanushek and Raymond (2005), in their analysis of national

US data, concluded that there is no clear evidence of this form of gaming in order to raise examination grades. Ehren and Swanborn (2012) found some evidence of reshaping the test pool in the Netherlands, but as there were no key differences between high and low performing schools it is possible this was due to motives other than the improvement of inspection evaluations.

Very little empirical research exists on unintended strategic behaviour. Sturman (2003) studied survey data from primary schools in England and found some evidence on teaching to the test. However, Sturman argues that these practices can have beneficial effects as well as negative effects leading to score inflation. Klein et al. (2000) and Tymms (2004) both suggest that test score rises, in Texas and English schools respectively, were partly due to teaching to the test practices. Wiggins and Tymms (2002) use survey data to compare primary schools in England (where league tables of examination results are published) with Scotland (where no results are published). They found large differences between Scottish and English schools, with English schools reporting more concentration on performance targets at the expense of other important objectives, a greater 'narrowing effect' on the curriculum and a greater focus on 'borderline' students (those close to the border for national target levels) at the expense of other students. Finally, Rosenthal (2004) measured a lowering of examination results in years in which schools had undergone inspection and hypothesised that this was due to time and energy being taken away from the students and focussed instead on the inspection.

Agreeing with De Wolf and Janssens (2007) and Penninckx (2016), we can see from this summary that there is limited evidence of 'teaching to inspection', 'sub-optimisation' and 'isomorphism' in the small number of studies which have been conducted in England, the United States and the Netherlands.

Profiling accountability systems in participating countries

An overview of the inspection system and wider accountability system in each country is presented below (for further details see Ehren et al. 2013). These descriptions refer to the systems as of 2011. There have been more recent changes in a number of countries, for example Austria no longer have inspections.

Austria

Since the early 1990s, Austria has undergone several phases of modernising the governance of the school system (Eder and Altrichter 2009). Largely as a consequence of international large scale comparative assessment studies, Austria has moved towards output-orientated methods and introduced – like many other European countries – an evidence-based governance regime (Altrichter and Heinrich 2007). This had led to educational standards, regular national assessment of student performance, partially-centralised final exams at the end of upper secondary schools, continuous system monitoring and national education reports have been established. School inspections, however, have only been introduced in some federal states (e.g. Styria) and meanwhile replaced by a "new quality management" for all schools on primary and secondary level (Kemethofer and Altrichter 2015). All monitoring

and accountability measures can be classified as low-stakes since they carry little in the way of pressure or sanctions (Specht and Sobanski 2012).

The Czech Republic

The external control of schools by central inspectorates has a long tradition in the Czech lands dating back to the 18th century, with the first centrally organised institution for the control of schools being founded in 1759. The current structure and mission of the Czech school inspectorate was established in 1995, when the main aim of inspection was reoriented towards the evaluation of the school as a whole, rather than the inspection of individual teachers, as was the case during the Communist regime. The Czech Republic has no national testing (except a final upper-secondary leaving examination introduced in 2011) and therefore inspection represents the main accountability mechanism and external control over schools which are, in comparative terms, highly autonomous (Greger and Walterová 2007). The main pressure mechanism exerted on schools comes from making inspection reports publicly available. In theory the inspectorate has the right to recommend to the ministry of education the closure of a school, but in reality this right is not used. Instead the pressure comes from the local education authorities that hire and promote school principals and influence their remuneration.

England

An Act of Parliament in England in 1992 made it compulsory for all state funded schools to be inspected. The same Act led to the creation of the Office for Standards in Education (Ofsted), the organisation responsible for inspections in England. Prior to this, there was a long history of inspections and accountability in England, commencing in the 1830s (Brighouse 1995). As far back as the 1860s there was a system of 'payment by results' where schools were given a large portion of their grants based on student performance (Wilcox and Gray 1996, 24). Due to this history, some scholars consider England to have 'one of the strongest accountability systems in the English speaking world' (Southworth 2002, pp192 in Barzano 2009). In addition to inspections, compulsory public examinations and published school league tables form an integral part of the accountability system in England today. This has led to a high-stakes system that puts pressure on schools to do well in order to avoid consequences such as the loss of pupils to other schools, extra monitoring inspections or in extreme cases school closure.

Ireland

The evaluation of schools by a centralized inspectorate has been part of Irish educational provision since the 19th century. Following the significant breakdown of the system in the 1990's (see for example Egan (2007) and McNamara and O'Hara (2008)) there has been a substantial refocusing of inspection beginning with the introduction of a cyclical approach to inspection called 'Whole School Evaluation' (McNamara et al. 2011). A move towards a more focused and intensive model then resulted in the introduction of a significantly strengthened inspection regime in 2012-2013 (O'Brien et al. 2014). This included the continued inspection of individual teachers but added requirements for schools to formally conduct self-evaluation and enabled the inspectorate to follow up schools deemed to be

underperforming. However, whilst there are national examinations, the compilation of school league tables is outlawed and there is no system of sanctions or rewards related to either school or teacher performance. The publication of inspection reports and the influence and perhaps prestige of the inspectorate represent the main elements of pressure to improve in the inspection system.

The Netherlands

During the last 20 years there has been a trend towards increased school autonomy and local decision-making in Dutch educational governance, however, more centralised arrangements have been implemented in the area of accountability. One central element of accountability measures is performance standards with centralised national testing (Scheerens et al. 2012), with the Dutch Inspectorate of Education having increased emphasis on the test and assessment results as an indicator of a school's performance (Béguin and Ehren 2010). A risk-based inspection system, including early warning analyses, was introduced in 2007 (Ehren and Honingh 2011). This makes use of student achievement results on standardised tests, self-evaluation and financial reports, parent complaints, and media coverage to decide whether a school is labelled as potentially failing and thus whether it will receive an inspection (Ehren and Swanborn 2012, 264). This inspection approach increases the pressure on the principals and teachers that are evaluated and, as a consequence, can generate strategic behaviour on the tests and lead to invalid assumptions when assessing the performance of a school (Béguin and Ehren 2010).

Sweden

After a long period of limited resources for the inspection of schools, a new authority for school inspection (the Swedish Schools Inspectorate or SSI) came into being in 2008. The Education Act of 2011 gives the inspectorate far-reaching powers. It can implement sanctions against municipal schools and withdraw public funding for independent schools that do not follow regulations. Since 2010 the SSI has adopted a system of risk-based inspection with more thorough inspections for schools judged as being at risk. In addition to regular inspection, thematic inspections are undertaken. These focus on particular areas of school activities (e.g. leadership) or the teaching of certain subjects. Parents have online access to written reports from school inspections as well as data on school performance in national tests adding pressure into the system for schools to do well.

Switzerland

In Switzerland the responsibility for the school system lies with the 26 cantons. Each canton has developed a public school system and established school laws. The cantons vary according to size and number of schools and governance structures (Huber 2011). 18 out of 21 German speaking cantons have so far implemented school inspection. The models differ on a range of aspects such as the organization, design, procedures, reporting and use of evaluation but it is accepted that all of them can be characterised as 'soft' as not much pressure is experienced by school leaders. At the same time, an emphasis on communication with schools and a high degree of transparency are typical across cantons (Huber 2011).

The above is a brief summation of very substantial accounts of school evaluation systems in the research partner countries. Based on these accounts a typology of the varying inspection approaches was developed including the level of pressure experienced by schools as a result of inspection. In this typology England and the Netherlands were described as high pressure inspection systems, Sweden, Ireland and the Czech Republic had medium pressure systems and Switzerland and Austria had the lowest pressure inspection systems. Items were then designed for inclusion in the survey to test whether unintended consequences of inspection could be identified and if so whether they were connected to the level of pressure applied.

Research method

Samples

The samples for this project were chosen to maximise the opportunity to make causal claims about the impact of school inspection. Randomised control trials were not an option due to the legal requirement for inspections in a number of countries. Instead, where possible (in the Netherlands and England) a regression discontinuity design (RDD) was used (see Kyriakides and Luyten 2009) requiring samples to be chosen from either side of particular thresholds. In the remaining countries, samples were chosen for a time series design. Surveys were sent to a minimum of 150 primary schools and 170 secondary schools in each country. More were sampled in some countries according to response rates and resources.

In England 211 primary schools and 211 secondary schools were chosen that scored closest to the *satisfactory* grade threshold in their main inspection 2009/10. An additional random sample was selected in order to assess whether any bias existed due to different sampling methods and to increase the number of respondents. This random sample consisted of 1246 primary schools and 637 secondary schools across all inspection grades from the same year. 246 schools were common to both samples. In the Netherlands three threshold groups were chosen (no risk, risk, high risk) each of which consisted of 100 schools (50 schools in primary education, 50 schools in secondary education), adding up to a selection of 300 schools in total. In order to allow for non-response of schools a targeted sample of 408 primary schools and 359 secondary schools was selected. In the remaining countries samples were chosen for a time series design. In Ireland and Austria (where data was collected from the federal state of Styria) all schools were selected for their sample (3200 primary schools and 729 secondary schools in Ireland and 503 primary and 194 secondary schools in Austria). In Sweden a random sample of 1167 primary and 987 secondary schools was selected and in the Czech Republic schools were targeted using the TIMSS design to sample schools (150 primary and 170 secondary). In Switzerland all primary and secondary (465) schools in 5 cantons were selected. The cantons were selected according to a typology developed after document analysis of school inspection systems in all German speaking cantons in order to get a broad spread of school inspection organisations.

[Insert Table 1 about here]

Table 1 presents the targeted sample sizes for each country and the actual number of participants that responded to the surveys, along with their response rates. Response rates varied greatly across countries, ranging from just 4% of primary schools in Ireland to 77% of secondary schools in Austria.

Survey instrument

A survey instrument for school principals was developed for the project and included items aimed at measuring the unintended effects of school inspections using a 5 point-Likert answer scale. Based on the identified categories in the literature of unintended consequences of school inspections (Fitz-Gibbon, 1997; De Wolf and Janssens, 2007), we included items on the extent to which school inspections lead to a narrowing of curriculum and instructional processes in the school (*convergence*), the extent to which principals experience inspections as an administrative burden (*formalisation* and *proceduralisation*) and the extent to which principals manipulate documents and data they send to the inspectorate (*misrepresentation*). Table 2 gives an overview of the items that were developed. The items did not form a scale, leading us to analyse unintended consequences by item. This had the benefit of allowing us to provide evidence on the broad variety of unintended effects of school inspection.

Method

Following a piloting phase the finalised survey was administered to school principals in the schools in each of the samples in the seven participating countries. The survey was repeated to the same sample three times, in 2011, 2012 and 2013. For this analysis, we have selected data from Year 1 of the project.

Descriptive statistics for each item are presented. The data are split by country and the countries are ordered by how much pressure the system creates (Altrichter and Kemethofer, 2014). Observations are made, and linked to interview data/open response data. To investigate our research questions, we use a multivariate analysis of variance (MANOVA). We report Eta^2 to describe the amount of variance explained in the dependent variables and we use Cohen's f^2 as an effect size (Cohen 1988, 477-478).

Limitations, threats and mitigation

We identify the following threats to the validity of our results and we present evidence to address these threats.

Risk of bias from sampling

The study uses two sample styles; some countries use the RDD sample selecting schools around a point on a continuum of school quality, others use random samples across all schools. To give confidence that the sampling styles are not introducing bias a separate random sample was selected in England alongside the RDD sample. T-tests were performed to compare the means of this RDD and random sample for all scales in the project. The standardised mean differences were small and there were no statistically significant differences between the two samples for any of the scales. This gives us confidence that,

where it can be tested, the different sampling does not appear to be introducing significant bias in the results.

Risk of bias from low response rates

Table 1 shows that there were low response rates in a number of countries so we must be cautious claiming any findings are representative. However we can strengthen our claims if we can show that the schools that responded are significantly different to schools that did respond. We have therefore, where possible, compared schools that responded with those that did not. Using national datasets (primary and secondary) in England (where data is more readily available compared to other countries in the study), schools responding to the survey were compared with those that did not on a number of characteristics (achievement in national examinations, value-added/progress measure in English and mathematics, a measure of socio-economic status, inspection ratings, proportion of students with special educational needs and proportion of unauthorized absences). The standardized mean differences were small with no statistically significant differences for either primary or secondary schools. In Austria there were high response rates and the schools in the sample do not differ significantly from the total population with reference to school type and size. Reasonable response rates were achieved in the Czech Republic and Sweden, and good regional coverage across the country and representation of schools with different size/student intakes was confirmed in the Czech Republic and good representation of school types in Sweden.

Risk of bias from missing data

Table 2 presents the number of principals that completed the unintended consequences items. Fewer respondents completed items directly related to inspection because not all schools in the time series design received an inspection in Year 1 of the study and in those schools that had received an inspection, not all staff were present during the inspection. The second column in Table 2 presents data on the number of respondents that were given items on inspection and column three presents the number of principals that answered all five of the unintended consequences items. The last column presents missing data for these items, which we can see ranges from 0 to 10%, which is considered small in terms of missing data, and should therefore limit any bias.

[Insert Table 2 about here]

Risk of bias from item wording and interpretation

There may be issues in interpreting our results, particularly across countries, due to issues of interpretation and translation. Firstly, differences may have been introduced when the surveys were translated from English into Dutch, Swedish, Czech and German. The items that were translated into Dutch from English were back-translated into English. The back translation matches the original English, giving us some confidence that meaning was not lost in translation. Similarly the Austrian and Swiss research teams cooperated to ensure a high quality of translation using back translation. Secondly, items may be interpreted differently by different people, within a country and across countries. We minimised this issue by interviewing convenience samples in each country during the pilot stage to check

respondents' understanding of items. Although the wording of items is a little clunky in places there was a common understanding of items both within countries and across countries.

Risk of bias from principals not responding honestly
This risk particularly applies to items that could be perceived to reveal negative behaviour (although the issue was not raised as a concern in the piloting phase). A separate part of the study collected data from teachers in England and the Netherlands and item 1 (see Table 3) was common to both surveys. The principals responded to whether or not they *discourage teachers to experiment with new teaching methods that do not fit the scoring rubric of the inspectorate* and teachers were asked whether or not *my principal discourages me to experiment with new teaching methods that do not fit the inspection criteria of the inspectorate.* If the principals were not answering honestly then we would expect to see a marked difference between the responses of the principals and teachers. However when principals and teachers were matched by school 80% of the responses matched within +/-1 Likert response, giving some evidence that in general principals are answering honestly.

Survey Results
From Table 2 we can see that across the seven countries, 1122 school principals replied to all items referring to *unintended consequences*. Table 3 presents the mean and standard deviation for each item. Figure 1 presents these results on charts with 95% confidence intervals. The countries are ordered according to how much pressure the inspection system creates, with the most highly pressurised system first.

[Insert Table 3 about here]

A multivariate analysis of variance (MANOVA) shows significant differences in unintended consequences across countries, with significant differences between countries being found with respect to every item (see MANOVA results below each chart in Figure 1). The ratio of variance explained by country varies between 4 % (presenting a positive picture) and 19% (narrowing curriculum) among the five items. According to Cohen (1988), f^2-values of 0.02 represent a small effect, values of 0.15 represent a medium effect and values of 0.35 represent a large effect. We can see from Figure 1 that all but item 4 show medium sized effects. A Tukey-test was used to identify which countries differ on a level of $p \leq .05$. Even when taking the 95% confidence intervals into account, there is a clear gap between many of the means of the countries, which shows that the countries differ in their prevalence of unintended consequences. We now go on to describe these differences.

[Insert Figure 1 about here]

Figure 1 indicates that the clearest patterns can be seen for items 2 and 3. The results from item 2 show an association between increasing pressure in a school inspection system (in terms of pressure to do well in an inspection) and an increase in the *narrowing* of the

curriculum and instructional strategies in the school. Similarly, item 3 shows that as the inspection systems increase in pressure there is an associated increase in principals claiming that school inspections have resulted in a *refocussing* of the curriculum and instructional strategies. Table 3 shows that in the high pressure systems of England and the Netherlands approximately 50% of the principals believe that inspection is resulting in a narrowing and refocussing of the curriculum and instructional strategies, whereas in low pressure systems like Switzerland and Austria this proportion is 10% or lower.

Based on the literature we would expect to see principals discouraging teachers from experimenting with new teaching methods that do not fit the scoring rubric of the inspectorate. The results from item 1 in Figure 1 show that the majority of principals claim they do not discourage teachers in this way and the level of pressure of the inspectorate does not appear to be related to their responses. However Table 3 shows that although it is not the norm to discourage teachers from experimenting there are a large number of principals who admit to this behaviour. In fact, on average across the countries surveyed one in ten principals attempt to curb their teachers' teaching practices to remain in line with what is perceived to be the inspectorate's preferences. Between 4% and 34% of principals answered positively to the item across the countries.

Similarly the literature would lead us to expect that inspections would lead a number of principals to misrepresent their school in the data sent to the inspectorate in order to make their schools look better. The results for item 4 showed the majority of schools do not claim to be painting a more positive picture of their school than actually exists in the data presented to school inspectorates. However, Table 3 shows that on average across the countries 7% of principals admit to presenting the school in a more positive light, i.e. misrepresenting their school to the inspectorate. This is particularly high in the Netherlands, England and Ireland where one in ten or more principals are admitting to misrepresenting their data to the inspectorate. Although this practice was not the norm, a result of 7% is high for such undesirable behaviour. By comparing means there seems to be no association between these proportions and how much pressure the inspection system creates, however it is noticeable that three of the four highest pressurised systems have the highest proportions of principals responding positively to misrepresent data.

Principals responded more strongly to item 5, with between 37% and 84% agreeing that the preparation for inspections is mainly about putting protocols and procedures in writing and gathering documents and data. The negative aspects of this are illustrated by an open response comment from a principal in England:

> *"Data drives everything and appears to be 90% of the inspection process, ignoring the wider holistic role of school."*
>
> [School Principal, England]

The results for item 5 do not show any association with pressure, with the medium pressurised systems of the Czech Republic and Sweden showing the greatest prevalence. This

finding for the Czech Republic probably reflects the mainaim of inspection in that country which is to ensure that the schools comply with formal regulations and legislation. Inspection in that country is therefore compliance rather than improvement-driven.

Discussion and summary

This project has given us a rare opportunity to empirically study some of the unintended outcomes of school inspection in a systematic way, analysing various consequences across a number of different countries. In all research we need to be cautious in our claims due to the threats to the validity of our results. However we have tested potential threats and bias in number of possible ways and have found no clear bias, giving us increased confidence in our findings. Findings from the survey showed varying prevalence of each unintended effect and varying association with pressure and it is clear that in some key areas there is a significant link between inspection pressure and negative unintended consequences.

Most clearly and most importantly, the results showed a clear association between increasing pressure in a school inspection system and an increase in the *narrowing* of the curriculum and instructional strategies in the school. Similarly, as the inspection systems increase in pressure there is an associated increase in principals claiming that school inspections have resulted in a *refocussing* of the curriculum and instructional strategies. In an era when the development of skills and attributes including creativity, problem-solving, self-directed learning and flexibility are considered vital educational outcomes this must be of serious concern.

Other indications of negative consequences of inspection also related to high pressure systems are less prevalent and yet very concerning. These include discouragement of experimenting with new teaching methods and the misrepresentation of the school in the data sent to the inspectorate. Even small numbers admitting to these actions is undesirable. Again, as with narrowing the curriculum, limiting teaching methods and teacher autonomy may also impact on the achievement of key educational objectives.

We found that some unintended consequences do not seem to be linked to the amount of pressure an inspection system creates. The reason for this is unclear. For example the prevalence of formalisation and proceduralisation is evident across all countries studied and is also a matter of concern. It is probable that coping with the demands of inspection systems places a heavy burden on school management and teachers and results in a great deal of time being spent on record-keeping, form-filling and related activity regardless of whether the inspection system is high-stakes or not. This burden, moreover, is likely to increase as most inspection systems now require schools to engage in systematic self-evaluation as part of accountability procedures.

This study has shown that pressure is associated with some unintended effects of school inspection, but did not find evidence that it is linked with others. Whilst the association does not establish causation, the findings are consistent with increased pressure producing unintended and negative consequences. This corresponds to previous empirical work, to Le

Chatelier's principle and Campbell's Law (1976) and to the predictions of key writers such as Smith (1995) and Fitz-Gibbon (1997).

References

Altrichter, H. and Kemethofer, D. (2015) Does Accountability Pressure through School Inspections Promote School Improvement? *School Effectiveness and School Improvement* 26 (1), 32-56.

Altrichter, H. and Heinrich, M. (2007) Kategorien der Governance-Analyse und Transformation der Systemsteuerung in Österreich. [Categories of Governance-Analysis and Transformation of System Governance in Austria] In *Educational Governance – Handlungskoordination und Steuerung im Bildungssystem [Educational Governance - Coordination of Action and Governance in the Educational System]* edited by H. Altrichter, T. Brüsemeister and J. Wissinger, 55-103. Wiesbaden: Verlag für Sozialwissenschaften.

Altrichter, H. Kemethofer, D. and Leitgöb, H. (2012) Ansätze der Systemsteuerung in der Einschätzung von Schulleitern. [How School Principals Perceive and Evaluate Strategies of System Governance] *Empirische Pädagogik*, 26 (1), 12-32.

Barzano, G. (2009) Reflecting on English educational accountability. *Italian Journal of Sociology of Education*, 3, 189-209.

Béguin, A., and Ehren, M. (2010) Aspects of accountability and assessment in the Netherlands. *Zeitschrift für Erziehungswissenschaft*, 14, 25-36.

Brighouse, T. (1995) The history of inspection In *School Inspection*, edited by. T. Brighouse and B.Moon, 1-14. (London, Pitman).

Brimblecombe , N.,Ormston, M. and Shaw, M. (1996) Teachers' perceptions of inspections, in: J. Ouston, P. Earleyand B. Fidler (Eds) *Ofsted Inspections: the early experience* (London, David Fulton Publishers).

Campbell, D.T. (1976) Assessing the Impact of Planned Social Change Occasional Paper Series, Paper 8, The Public Affairs Center, Dartmouth College, Hanover New Hampshire, USA. December, 1976

Case, P., Case, S. and 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.

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

Cohen, J. (1988) *Statistical Power Analysis for the Behavioural Sciences.* (Lawrence Erlbaum Associates, Hillsdale) 477-478

Cullen, J B and Reback, R. (2006) Tinkering toward accolades: school gaming under a performance accountability system, in *Advances in Applied Microeconomics Vol.14 Improving School Accountability: Check-Ups or Choice,*. Edited by T. Gronberg and D. Jansen, (Amsterdam, Elsevier Science) 1-34.

De Wolf, I. F. and Janssens, J. G. (2007) Effects and side effects of inspections and accountability in education: an overview of empirical studies. *Oxford Review of Education,* 33(3), 379–396.

Eder, F. and Altrichter, H. (2009) Qualitätsentwicklung und Qualitätssicherung im österreichischen Schulwesen: Bilanz aus 15 Jahren Diskussion und Entwicklungsperspektiven für die Zukunft. In: *Nationaler Bildungsbericht Östereich 2009. Band 2: Fokussierte Analysen bildungspolitischer Schwerpunktthemen* edited by W. Specht, 305-322. (Graz, Leykam).

Egan, E. (2007) The evaluation of teachers and teaching in primary and post primary schools by the inspectorate of the department of education and science. In *The competences approach to teacher professional development: current practice and future prospects* edited by R. Dolan and J. Gleeson, 37-39. (Armagh, The Centre for Cross Border Studies).

Ehren, M. C. M., Altrichter, H., McNamara, G. and O'Hara, J. (2013) Impact of school inspections on improvement of schools – describing assumptions on causal mechanisms in six European countries *Educational Assessment, Evaluation and Accountability* 25 (1), 3-43.

Ehren, M.C.M., and Honingh, M.E. (2011) Risk-based school inspections in the Netherlands: A critical reflection on intended effects and causual mechanisms. *Studies in Educational Evaluation* 37, 239-248.

Ehren, M.C.M. and Swanborn, M.S.L. (2012) Strategic data use of schools in accountability systems. *School Effectiveness and School Improvement* 23(2), 257-280.

Ehren, M.C.M. (2006) Toezicht en schoolverbetering [School inspection and school improvement] PhD diss., (Delft, Eburon).

Eurydice (2004) *Evaluation of Schools providing Compulsory Education in Europe*. Available online at: eacea.ec.europa.eu/education/eurydice/documents/all_publications.pdf (accessed 6th June 2014).

Figlio, D.N. and Getzler, L.S. (2006) Accountability, ability and disability: gaming the system, in T. Gronberg and D. Jansen (Eds) *Advances in Applied Microeconomics Vol.14 Improving School Accountability: Check-Ups or Choice,* (Amsterdam, Elsevier Science) 14, 35-49.

Fitz-Gibbon, C. T. (1997) *The Value Added National Project: Final Report: Feasibility studies for a national system of Value Added indicators*. (London, School Curriculum and Assessment Authority).

Fitz-Gibbon, C.T. and Stephenson-Forster, N.J. (1999) Is Ofsted helpful? An evaluation using social science criteria, in C. Cullingford (Ed) *An Inspector Calls: Ofsted and its effect on school standards.* (London, Kogan Page) 97-118.

Goodhart, C. (1975) reprinted in Goodhart, Charles (1981). Problems of Monetary anagement: The U.K. Experience, in Anthony S. Courakis (Ed.), *Inflation, Depression, and Economic Policy in the West* (Rowman & Littlefield) 111–146

Greger, D. and Walterová, E. (2007) In Pursuit of Educational Change: The Transformation of Education in the Czech Republic. *Orbis scholae* 1(2), 11–44.

Hanushek, E. A., and Raymond, M. E. (2005) Does School Accountability Lead to Improved Student Performance? *Journal of Policy Analysis and Management* 24(2), 297-327.

Huber, S. G. (2011) School Governance in Switzerland: Tensions between New Roles and Old Traditions. *Educational Management Administration & Leadership, 39*(4), 469-485.

Jacob, B. A. (2005) Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago public schools *Journal of Public Economics* 89(5), 761–796.

Jacob, B. A. and Levitt, S. D. (2003) Rotten apples, an investigation of the prevalence and predictors of teacher cheating *The Quarterly Journal of Economics* 118(3), 843–877.

Jones, K. and 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.

Kemethofer, D. & Altrichter, H. (2015). Schulqualität Allgemeinbildung (SQA) in der Einschätzung von Schulleitungen allgemeinbildender Pflichtschulen. Erziehung und Unterricht, 165(7-8), 675-690.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F. and Stecher, B. M. (2000) *What do test scores in Texas tell us?* (Santa Monica, CA, Rand).

Kyriakides, L. and Luyten, H. (2009) The contribution of schooling to the cognitive development of secondary education students in Cyprus: an application of regression-discontinuity with multiple cut-off points. *School Effectiveness and School Improvement* 20(2), 167-186.

Leeuw, F. L. (2000) Onbedoelde neveneffecten van outputsturing, controle en toezicht? In Raad voor Maatschappelijke Ontwikkeling, Aansprekend burgerschap; de relatie tussen de organisatie van het publieke domein en de

verantwoordelijkheid van de burgers (Den Haag, Raad voor Maatschappelijke Ontwikkeling), 151–171. [Unintentional side effects of output manipulation, control and school inspections?, In Responsible citizenship; the relationship between the organisation of the public domain and the responsibility of the citizens].

McNamara, G and O Hara, J. (2008) *Trusting Schools and Teachers: Developing Educational Professionalism Through Self-Evaluation* (New York, Peter Lang Publishing Inc.).

McNamara, G., O'Hara,J., Lisi, P., and Davidsdottir, S. (2011) Operationalising self-evaluation in schools: experiences from Ireland and Iceland. *Irish Educational Studies*, 30(1), 63-82.

Merton, R. K. (1936) The unanticipated consequences of purposive social action. *American Sociological Review* 1, 894-904.

O'Brien,S. , McNamara, G and O'Hara, J. (2014) Critical Facilitators: External Supports for Self-Evaluation and Improvement in Schools *Studies in Educational* Evaluation 43, 169-177

Penninckx, M. (2016). Effects and side effects of school inspections: A general framework. *Studies in Educational Evaluation*. DOI: 10.1016/j.stueduc.2016.06.006

Perryman, J. (2007) Inspection and emotion. *Cambridge Journal of Education*, 37(2), 173-190.

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

Scheerens, J., Ehren, M., Sleegers, P., and de Leeuw, R. (2012) *OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes. Country Background Report for the Netherlands*. (OECD Publishing) Available online at: http://www.oecd.org/edu/school/NLD_CBR_Evaluation_and_Assessment.pdf (Accessed 5 January 2016)

Southworth, G. (2002) School Leadership in English Schools, in Walker, A. and Dimmock, C. (Eds) *School Leadership and Administration: adopting a cultural perspective* (London: RoutledgeFalmer) 187 – 204.

Specht, W. and Sobanski, F. (2012) OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes. Country Background Report for Austria. (OECD Publishing) Available online at: http://www.oecd.org/edu/school/49578470.pdf (Accessed 5 January 2016)

Smith, P. (1995) On the unintended consequences of publishing performance data in the public sector. *International Journal of Public Administratio*n 18(2 and 3), 277-310.

Sturman, L. (2003) Teaching to the test: science or intuition? *Educational Research* 45(3), 261–273.

Thomsen, V. B. E. (2000) Le Châtalier's Principle in the Sciences *Jounal of Chemical Education* 77 (2), 173-176

Tymms, P. (2004) Are standards rising in English primary schools? *British Educational Research Journal* 30(4), 477-494.

Wiggins, A. and Tymms, P. (2002) Dysfunctional Effects of League Tables: A Comparison between English and Scottish Primary Schools. *Public Money and Management* 22(1), 43-48.

Wilcox, B. and Gray, J. (1996) Inspecting schools: holding schools to account and helping schools to improve Buckingham, Open University Press.

Table 1: Sample sizes and response rates for Year 1 data collection

Country	Targeted Sample		Actual sample - Year 1 (response rate in brackets)				
	Primary	Secondary	Primary		Secondary		Combined
Netherlands	408	359	73	(18%)	15	(4%)	88
England*	1422	637	189	(13%)	101	(16%)	290
Sweden	1167	987	567	(49%)	464	(47%)	1031
Ireland	3200	729	123	(4%)	42	(6%)	165
Austria (Styria)	503	194	345	(68%)	149	(77%)	494
Czech republic	150	170	56	(37%)	69	(41%)	125
Switzerland				132			132
Total							**2325**

* Combined RDD and random samples

Table 2: Respondent numbers for unintended consequences items and missing data

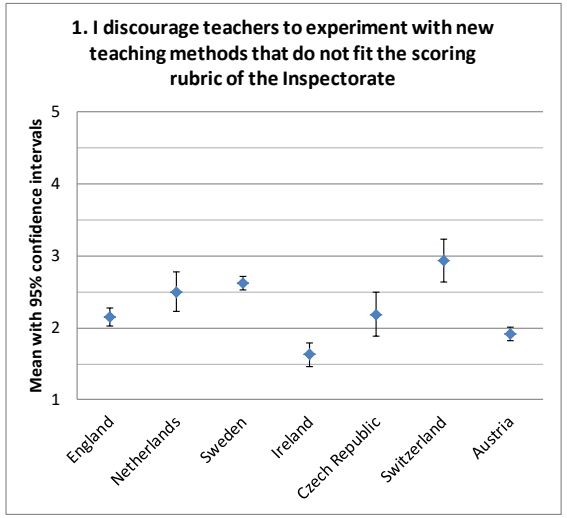
Countries	Respondents inspected/asked items 1-5	Number completed all 5 items	Percentage missing 1 or more items
Netherlands	47	45	4%
England*	238	229	4%
Sweden	362	343	5%
Ireland	128	114	9%
Austria (Styria)	320	288	10%
Czech republic	43	43	0%
Switzerland	65	60	8%
Total	**1203**	**1122**	

* Combined RDD and random samples

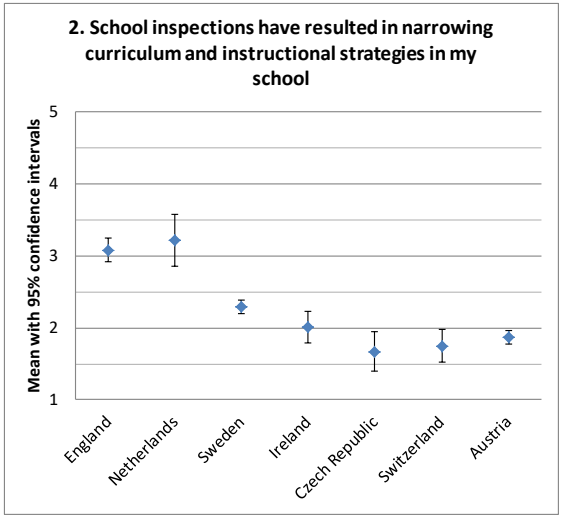
Table 3: Unintended consequences items and scales with descriptive statistics for combined Y1 and Y2 data

Item	Responses	Country	Number of schools	Mean	SD	% Positive response ("4" or "5")
1. I discourage teachers to experiment with new teaching methods that do not fit the scoring rubric of the Inspectorate	1 = strongly disagree 2 = disagree 3 = neither agree or disagree 4 = agree 5 = strongly agree	England	235	2.15	0.93	9
		Netherlands	46	2.50	0.94	15
		Sweden	354	2.62	0.87	12
		Ireland	123	1.63	0.94	4
		Czech Republic	43	2.32	1.01	7
		Switzerland	61	2.93	1.17	34
		Austria	302	1.91	0.79	4
		All countries	**1164**	**2.22**	**0.95**	**10**
2.School inspections have resulted in narrowing curriculum and instructional strategies in my school	1 = strongly disagree 2 = disagree 3 = neither agree or disagree 4 = agree 5 = strongly agree	England	238	3.08	1.25	40
		Netherlands	46	3.22	1.20	51
		Sweden	353	2.29	0.01	9
		Ireland	122	2.01	1.25	12
		Czech Republic	43	1.67	0.89	19
		Switzerland	61	1.75	0.91	5
		Austria	299	1.87	0.76	2
		All countries	**1162**	**2.36**	**1.11**	**15**
3. School inspections have resulted in refocusing curriculum and teaching and learning strategies in my school	1 = strongly disagree 2 = disagree 3 = neither agree or disagree 4 = agree 5 = strongly agree	England	238	3.53	1.06	60
		Netherlands	46	3.61	1.06	65
		Sweden	353	2.89	1.07	32
		Ireland	123	2.89	1.21	34
		Czech Republic	43	2.12	1.03	7
		Switzerland	61	2.03	1.06	10
		Austria	300	2.40	0.89	10
		All countries	**1164**	**2.92**	**1.13**	**31**
4. The latest documents/facts and figures we sent to the Inspectorate present a more positive picture of the quality of our school then how we are really doing	1 = strongly disagree 2 = disagree 3 = neither agree or disagree 4 = agree 5 = strongly agree	England	232	2.06	0.02	10
		Netherlands	46	2.02	1.13	15
		Sweden	350	1.93	0.77	2
		Ireland	122	2.23	1.18	14
		Czech Republic	43	1.58	0.85	2
		Switzerland	60	2.35	1.09	8
		Austria	292	2.23	0.88	7
		All countries	**1145**	**2.08**	**0.93**	**7**
5. Preparation for school inspection is mainly about putting protocols and procedures in writing that are in place in the school and gathering documents and data	1 = strongly disagree 2 = disagree 3 = neither agree or disagree 4 = agree 5 = strongly agree	England	232	3.12	1.24	46
		Netherlands	46	3.35	1.18	52
		Sweden	352	3.93	0.78	79
		Ireland	126	2.94	1.33	37
		Czech Republic	43	4.14	0.83	84
		Switzerland	60	3.65	1.15	70
		Austria	295	3.26	1.00	46
		All countries	**1154**	**3.46**	**1.11**	**58**

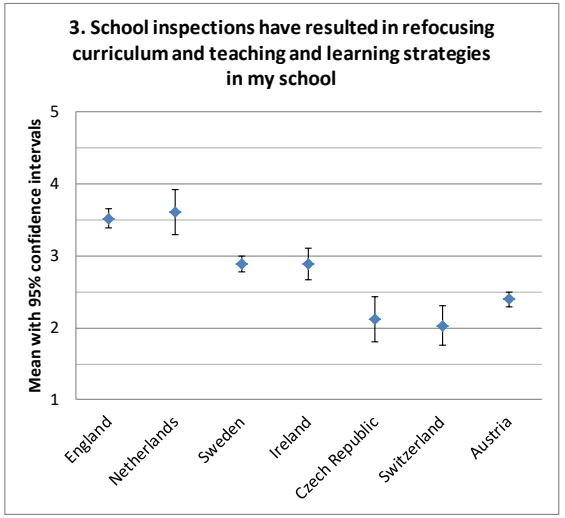
Figure 1: Charts showing responses to items, split by country and ordered by decreasing pressure



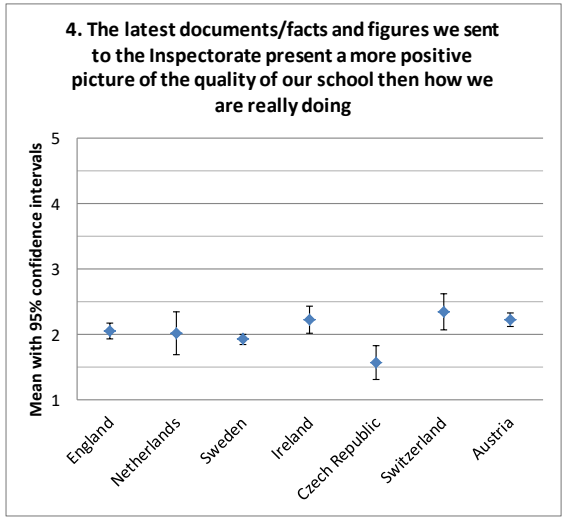
Difference between countries is sig. at $p<0.01$, $Eta^2=0.16$, $f^2=0.18$ (MANOVA)



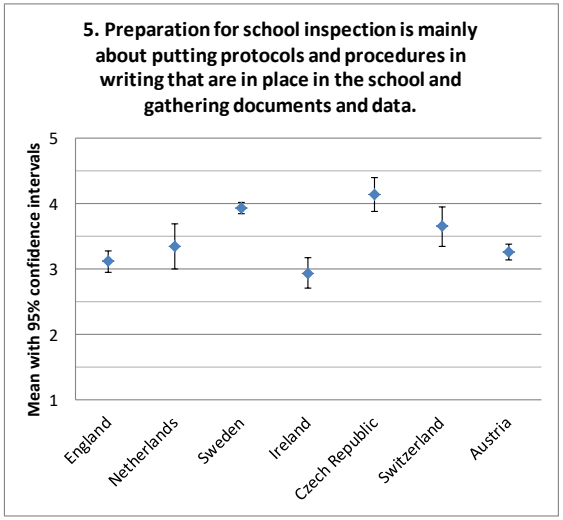
Difference between countries is sig. at $p<0.01$, $Eta^2=0.19$, $f^2=0.24$ (MANOVA)



Difference between countries is sig. at $p<0.01$, $Eta^2=0.17$, $f^2=0.19$ (MANOVA)



Difference between countries is sig. at $p<0.01$, $Eta^2=0.04$, $f^2=0.03$ (MANOVA)



Difference between countries is sig. at $p<0.01$, $Eta^2=0.13$, $f^2=0.14$ (MANOVA)

5 = Strongly agree

4 = Agree

3 = Neither agree or disagree

2 = Disagree

1 = Strongly disagree