Crimes Against Statistical Inference: Forcing Teachers to be Accessories After the (Absence of) Fact

Rosie Ridgway and Jim Ridgway
School of Education, Durham University
r.a.ridgway@durham.ac.uk and jim.ridgway@durham.ac.uk

Abstract

Reports on pupil performance form an important element in the efforts to improve the quality of education. Here, we examine the practicalities of making reliable judgements about changes in school performance over time. In a very large number of primary schools in England, too few pupils provide performance data to allow sensible conclusions to be drawn. In fact, the sample sizes are often not big enough to reliably detect the growth in attainment expected over a whole school year – let alone nuances about whether the performance by the same age group in adjacent years has changed. We provide links to simulations to illustrate this argument. Statistical inference is a difficult topic, and few teachers are able to draw conclusions from data that a professional statistician would endorse. We set out some heuristics developed by Wild and his co-workers (2011, in press) that can remedy this problem. We are critical of current practices and current plans to continue to perpetrate crimes against inference, and make suggestions about how this can be avoided. We highlight the strong negative impact that the poor statistical literacy embedded in government reporting requirements for schools can have on the lives of individuals, and on the reputation of schools.

Keywords

statistical inference, primary schools, reports, simulation, statistical literacy, league tables

Successive governments have set out to improve the quality of the education system in England in a variety of ways. One approach has been to establish a system where pupil performance is assessed at certain Key Stages via common tests, and to judge the performance of schools on the basis of the performance of relevant students. Tests have two distinct uses: one is for external consumption – league tables (attainment and achievement tables) are made available, so that informed consumers (parents and guardians) can make better choices about where to send their children; the other is to support reflection within the school about successes and failures, so as to inform future practice. Here, we explore the extent to which the evidence available to schools can support the weight of these interpretations, with a particular focus on primary schools. We describe the process of data collection and reporting data in schools, then examine sampling error in the context of small schools, using simulations to illustrate key points (over 40% of primary schools will be judged on the basis of the performance of 30 or fewer pupils). Methods for drawing robust conclusions from small samples are given, that can be used by people of varying levels of statistical sophistication. We conclude by exploring the implications for schools and individuals who are put in the invidious position of being evaluated on the basis of unreliable evidence.

How Tests are to be Used to Improve School Performance

The guidance for schools http://www.education.gov.uk/schools sets out a tiered system whereby teachers, subject leaders, school leaders, governors, School Improvement Partners (SIP) and Local Authorities monitor and analyse pupil performance data using tools such as those provided by the Reporting and Analysis for Improvement through School Self-Evaluation system (RAISEonline). These roles will be described briefly, as an introduction for those who are unfamiliar with the UK primary school context (see also the *Primary Schools Data Submission Guidance* (2010)).

Teacher: administers appropriate tests or teacher assessments (depending on year group), to track pupil progress, to report individual pupil progress to parents and to analyse data (e.g. by using the RAISEonline facility to examine value added scores). Teachers must also set targets for individual pupils in all year groups, update pupil tracking records, and analyse the performance data of individual pupils, thereby identifying under achieving groups or individuals.

Subject leader: identifies the implications from teacher assessments, makes use of RAISEonline to analyse data (e.g. question by question analysis of nationally mandated tests such as the Key Stage (KS) 1 and KS2 Standard Assessment Tests (SATs) – KS1 covers the age range up to Year 3, and KS2 to Year 6), and reports on subject results, results for groups of pupils, and implications for teaching and learning. The subject leader (SL) should take account of national results, collate individual targets for year groups, and Contextual Value Added (CVA) data (for KS1 to KS2). The SL should identify projected results and issues for teaching and learning from a scrutiny of work; the SL may identify requirements for professional development in specific subjects.

Leadership team: should keep the school's tracking record up to date (taking into account revised individual targets), and should communicate the information to governors. The teacher assessments or test results should be examined in terms of their implications for whole school teaching and learning. These should be collated into a whole school evaluation report (which should compare results of groups, year on year results, actual versus estimated results, and the proportion of pupils meeting expected levels). The report should be shared across the school.

Governors, SIP and the Senior Management Team: should propose targets for the end of KS2 and should aggregate results from individual targets for the current Year 5 pupils, and match these data against performance from previous years. They should identify draft statutory targets for Year 6 performance, and should identify the progress of groups, classes, and year groups - taking action where appropriate and drawing up a report for governors. The school's tracking system should be up to date and used to monitor a range of performance indicators, in particular progress by pupils. The school's governors should review evaluations, looking at evidence of progress and pupil performance, comparing performance against targets. They should also review attendance, and receive an annual report from the SIP.

Outside school: data are analysed by the SIP and Local Authority, as well as by the Office for Standards in Education (OFSTED).

These activities illustrate the continuous process of evaluating performance data within primary schools. The activities are important for schools because the data has great significance for the outcome of inspections by OFSTED (poor pupil performance is now a limiting judgement, which can downgrade a school's rating). Teachers, subject leaders, senior managers and Head teachers, are encouraged to improve performance and *get results* so that their schools can be judged as *good*.

The types of data required are currently under review, due to the change of government (2010). However, the current guidance for schools on the use, publication and significance of attainment data will remain unchanged. The types of school data which will be used in 2010 (from the April statement of intent 2010) will include:

- Contextual Value Added data (here, this is the estimated individual pupil progress from KS1 to KS2 with confidence intervals);
- Progress measures - the percentage of pupils making at least the expected level of progress in English and mathematics;
- Percentage of pupils achieving level 4 and above in mathematics and English in KS2 tests, (and teacher assessments);
- Percentage of pupils achieving level 5 and above in mathematics and English in KS2 tests, (and teacher assessments);
- Year on year comparisons: percentage of pupils achieving level 4 and above in English and mathematics tests in 2008, 2009, 2010 presented as a bar chart of aggregate percentages of test scores.

Each of these measures is defined clearly in documents for schools and parents (e.g.

http://www.education.gov.uk/performancetables/primary_09/pdf_09/840.pdf

How Good a Reflection of School Performance Can the Data Be?

The main purpose of schools is to educate children, and it is entirely appropriate that school performance should be examined, and made public. The first issue to be considered here is the soundness of the judgements that can be based on the data available.

Almost every sample taken from a population misrepresents that population to some extent. Two samples taken from the same population will rarely be the same, or have the same properties (e.g. identical ranges or medians). The big idea underpinning statistical inference is that it is possible to make a decision about populations that are based on samples from populations, and to have a good idea about how likely it is that the decision will be correct. Typical questions might be: are 12 year old girls taller, on average than 12 year old boys? Do UK students do better on an international test than American students? Do pupils who attend School A achieve higher scores than the national average? The extent to which a sample of a particular size represents its parent population can be explored by taking samples of the same size over and over again, and by looking at the variation in some measures – such as mean or interquartile range.

Wild, Pfannkuch, Regan, and Horton (2011, in press) provide an excellent pedagogic tool to inform conceptions of sampling variability. They pose the question 'are 14 year old girls taller on average than 13 year old girls?' The answer is obvious, and the size of the effect is quite large (judged relative to effects often associated with educational interventions). The Census at School project is an international project designed to promote statistical enquiry across the curriculum. In one component, pupils take an online survey that then provides data which can be accessed and analysed. Data collected include biometric measures (height, arm span etc.) and a range of topics of interest to pupils (such as access to different technologies, interest in environmental issues, bedtimes etc). The data from Census at Schools for New Zealand, based on measures from 5153 girls at age 13, shows a mean height of 161.8 cm (*sd*=10.45) while 4127 girls at age 14 show a mean height of 165.7 cm (*sd*=10.59). The effect size is 0.37.

This effect size can be put in the context of gains in educational attainment. 'An effect size of *d*=1.0 indicates an increase of one standard deviation on the outcome… a one standard deviation increase is typically associated with advancing children's achievement by two to three years.' Hattie (2009 p8). So every year, on average, the effect size for pupils' academic gain is roughly between 0.33 and 0.5 standard deviations. The effect size for differences in girls' height between ages 13 and 14 years is within this band, so we can use the question about girls' height differences as a surrogate for detecting educational progress, and can ask 'with sample sizes of 30, can we reliably detect a genuine gain of one year's educational progress?'

Wild and his co-workers have used simulations where samples of different sizes are selected randomly from these two populations, and are displayed as box plots (here, the boxplots shows the median, the median interquartile range, and the range). Initially, pairs of samples are drawn and displayed; one for 13 year olds, and one for 14 year olds. Later, the simulation repeatedly samples from the distributions, and each sample leaves a ghost of each box plot on screen. This is illustrated in Figure 1.

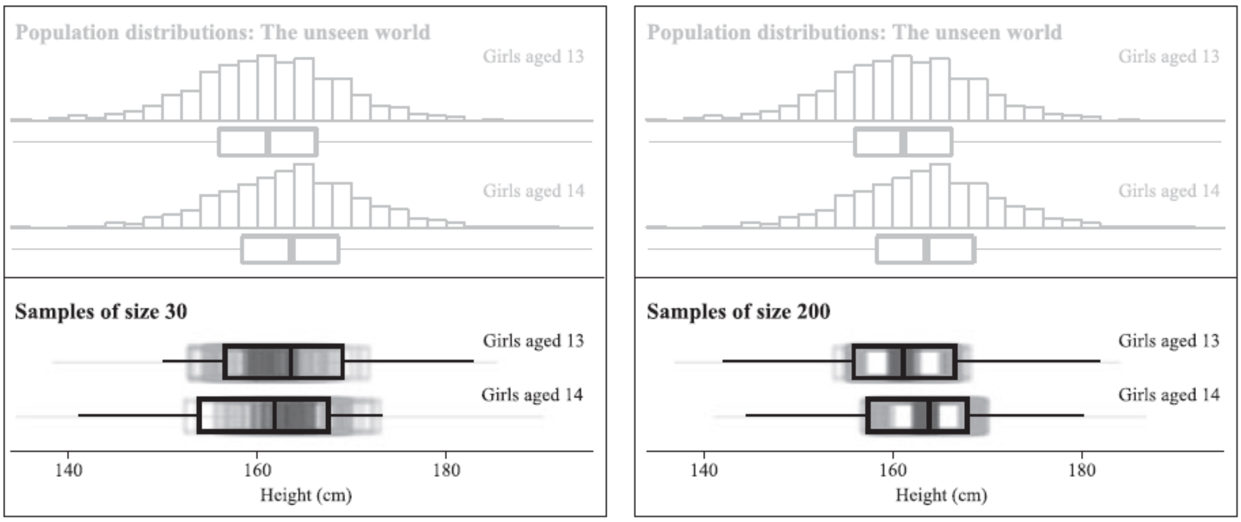

Figure 1. Repeated Sampling Box Plots: Plot with Ghosts from Earlier Samples

The display for samples of size 30 is contrasted with the display for samples of size 200. In the case of the smaller sample size, the median height of a sample of 13 year old girls will frequently be higher than the median height of 14 year old girls. This will rarely be the case when samples of size 200 are taken.

It is difficult to overstate the counter-intuitive nature of the results of repeated sampling with samples as small as 30 pupils. The reader is strongly encouraged to visit the simulation before continuing to read this text. http://www.censusatschool.org.nz/2009/informal-inference/WPRH/. We can now answer the question posed earlier: with sample sizes of 30, can we reliably detect a genuine gain of one year's educational progress? – Often, we will fail to do so. This result is important whenever samples sizes are small. In the case of small primary schools, the evidence is relevant to whole school issues. In the case of very large schools, the issue is still relevant, because of the reporting requirements that are based on small samples (such as the attainment of disadvantaged children). Teacher evaluations will almost always be based on samples not much (if at all) larger than 30. We return to the analysis of subpopulations in a subsequent section.

Wobbly Evidence in Context: Sizes of Primary Schools

Table 1 shows Department of Education data for January 2010 on the 16,971 primary schools in England in terms of school enrolment. We have added a calculation of the maximum cohort of pupils that will form the basis for judgements about the success of the school. We have assumed that each school comprises 7 age levels, and have calculated the maximum cohort size by dividing the maximum school size of each group by 7 (so for the group 101 to 200, we calculate 200/7). (This will produce a somewhat generous estimate of the maximum cohort size).

Distribution of Primary School Sizes in England									
	Up to 100	101 - 200	201 - 300	301 - 400	401 - 500	501 - 600	601 - 700	701 - 800	801 - over
Maximum Cohort	14	29	43	57	71	86	100	114	>114
	2,499	4,819	5,190	2,323	1,680	226	191	27	16

Table 1: Primary School Sizes in England

We conclude that more than 40% (i.e. 2,499+4,819 out of 16971) of all primary schools are being judged on cohorts of average size less than 30 pupils. The national picture will mask regional variations; in the Durham Local Education Authority, for example, 147 out of 211 schools had 30 or less pupils in the cohort aged 10 years. So small schools are required to monitor changes in attainment by the same year group (e.g. Year 6) in successive years, in a context where detecting differences between the attainment of Year 5 and Year 6 students is an uncertain business.

An illustration of the wobbliness of the data can be observed by looking at year-on-year performance in the league tables published for schools (http://www.education.gov.uk/performancetables/primary_09/pdf_09/840.pdf).

A further illustration will be provided. The measure *Contextual Value Added* (CVA) sets out to measure the progress made by pupils in their test results from the end of KS1 to the end of KS2. '*It takes into account the varying starting points of each pupil's KS1 test results, and also adjusts for factors which are outside a school's control (such as gender, mobility and levels of deprivation) that have been observed to impact on pupil results.*' Conceptual (what do you take into account?) and technical (how reliable is the measure) issues are beyond the scope of this paper, where we focus on far simpler problems.

No measure can be exact, so every measure should have some bounds around it, to give a feeling about its accuracy. Confidence intervals are given for CVA, and there is a clear statement that confidence intervals are smaller for big samples than for large ones. This is illustrated in Table 2 where the confidence intervals are given for three schools of different sizes (roughly 300, 200, and 100 pupils) with the same CVA, along with the locus of each school in a national league table of all maintained mainstream schools (unfortunately, the idea of a 95% confidence interval is not straightforward – the confidence level (here 95%) 'is the probability that the procedure that is used to determine the interval will provide an interval that includes the population parameter' (Utts and Heckard, 2007 p406)). We return to the problem of applying statistical inference appropriately without necessarily having deep statistical knowledge in a later section.

School	Enrolment	CVA (rank)	Upper CVA (rank)	Lower CVA (rank)
Dene House	293	100.0 (top 40%)	100.5 (top 40%)	99.5 (top 75%)
Neville's Cross	203	100.0 (top 40%)	100.7 (top 25%)	99.4 (top 75%)
St Joseph's	104	100.0 (top 40%)	101.0 (top 25%)	99.3 (top 95%)

Table 2. Bounds on CVA for Durham Schools of Different Sizes, with 'league table' Interpretations

The dependence of the confidence interval on school size is apparent. So too, is the awareness of the authors of the report concerning the problems of working with small data sets.

There are some major conceptual problems with the guidance given to teachers in a number of official documents. Errors of interpretation are set to continue if current plans for reporting school performance are adopted. A single example will be given here, taken from *A School Report Card: Prospectus* (DCSF, 2009).

'The new School Report Card, to be introduced from 2011, will provide our key statement on the outcomes we expect from schools, and the balance of priorities between them, ensuring more intelligent accountability across schools' full range of responsibilities.' P3

Table 3 shows part of a 'Scenario' taken from p39 of the *Prospectus* that explains the ways that schools will be given credit (or not) for reducing the gap between disadvantaged and other pupils. The text explains that School 1 has reduced the gap, so will be given credit, whist School 2 has not reduced the gap, so will not be given credit. The aim is laudable, but the conclusions drawn from the evidence presented here just cannot be justified.

% attaining Level 4 or above in both English and mathematics at KS2					
	Disadvantaged pupils		Peer pupils		Credit?
	2007	2008	2007	2008	
National	**51%**	**54%**	**75%**	**76%**	
School 1	45%	↑↑47%	76%	↑77%	Yes
School 2	45%	↑↑47%	76%	↑↑↑79%	No

Table 3. Table taken from *A School Report Card: Prospectus*

Exploration of the properties of samples of 300 selected from a population http://www.censusatschool.org.nz/2009/informal-inference/WPRH/ show the inherent instability of the properties of samples that are very large in comparison to those available in most schools. Later 'Scenarios' have column entries that include *insufficient pupils* showing that there is some awareness of the errors being made. We leave it to the reader to estimate the minimum number of pupils in each group that are required for the percentages in Table 3 to be given with such accuracy, and the implications for the sizes of the two schools in the example given here. Our conclusion is that it is quite possible that there is *not a single* primary school in England which is big enough to justify the sort of analysis suggested in Table 3. We conclude that this is not a good way of achieving the goal of *'ensuring more intelligent accountability'.*

Interpreting Data from Small Samples: A Primer for Primary Schools

Primary schools are required to submit data about, and analyse, the performance and progress of their pupils. One might expect that this would require an understanding of basic statistical principles, and an understanding of the complex nature of the data analyses required; however little or no professional development is provided to teachers. The training which is available (to subject leaders, head teachers, SIP) focuses on the production and input of data, and on identifying groups within performance data, rather than on the inferences that can be supported by the data available to the school. A key ambition for the Department for Education should be to provide tools that teachers can use to make judgements that would be endorsed by a professional statistician. A bonus would be to develop statistical intuitions and robust statistical inference. We need to develop ways to introduce core ideas in statistical inference that are simple to understand and act on.

The principles and practice of statistical inference are difficult to acquire (see Wild, Pfannkuch, Regan, and Horton (2010)) for a review of research on informal statistical inference), and it is unreasonable to expect that the majority of primary school teachers will bring this knowledge to their profession, or will acquire it on the job. This gives rise to an apparent impasse – teachers need to make decisions based on statistical inference, but do not have the skills, or the time to acquire the skills, to make these decisions.

Wild, Pfannkuch, Regan, and Horton (2011, in press) offer a solution to this apparent impasse, in the form of guidance on 'how to make the call' about plausible conclusions from samples of data. Their

Figure 8 (p13) is reproduced here as Figure 2. For some patterns, such as identical distributions, and non-overlapping interquartile ranges, sample size is irrelevant (subject to the warning that any inferences from sample sizes of 20 or less should be avoided); for intervening patterns, where the interquartile ranges overlap, sample size becomes important.

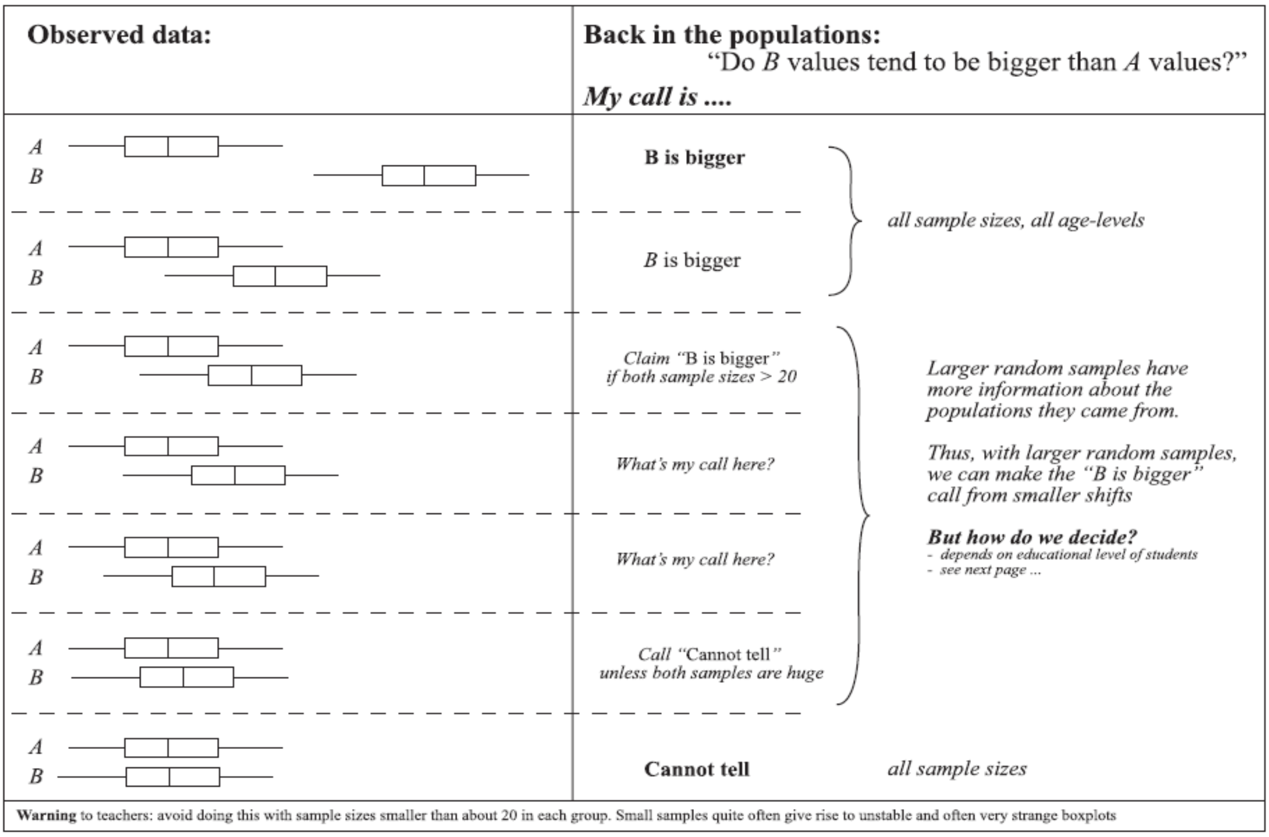

Figure 2: Decision Rules for Box Plots (from Wild et al, 2011, in press)

They provide specific guidance on inferences that can be made that are usable by people at all levels of statistical sophistication. Users' statistical sophistication can be documented via 'milestones' – these are set out in Figure 3 (which reproduces Wild et al's Figure 9). From the viewpoint of teachers deciding if differences are statistically reliable, all that is needed is a description of Milestones 1 and 2, and a look-up table to allow them to find the correct rule for the sample sizes they are dealing with. Teachers would key in their sample size(s) and the appropriate decision rule would appear.

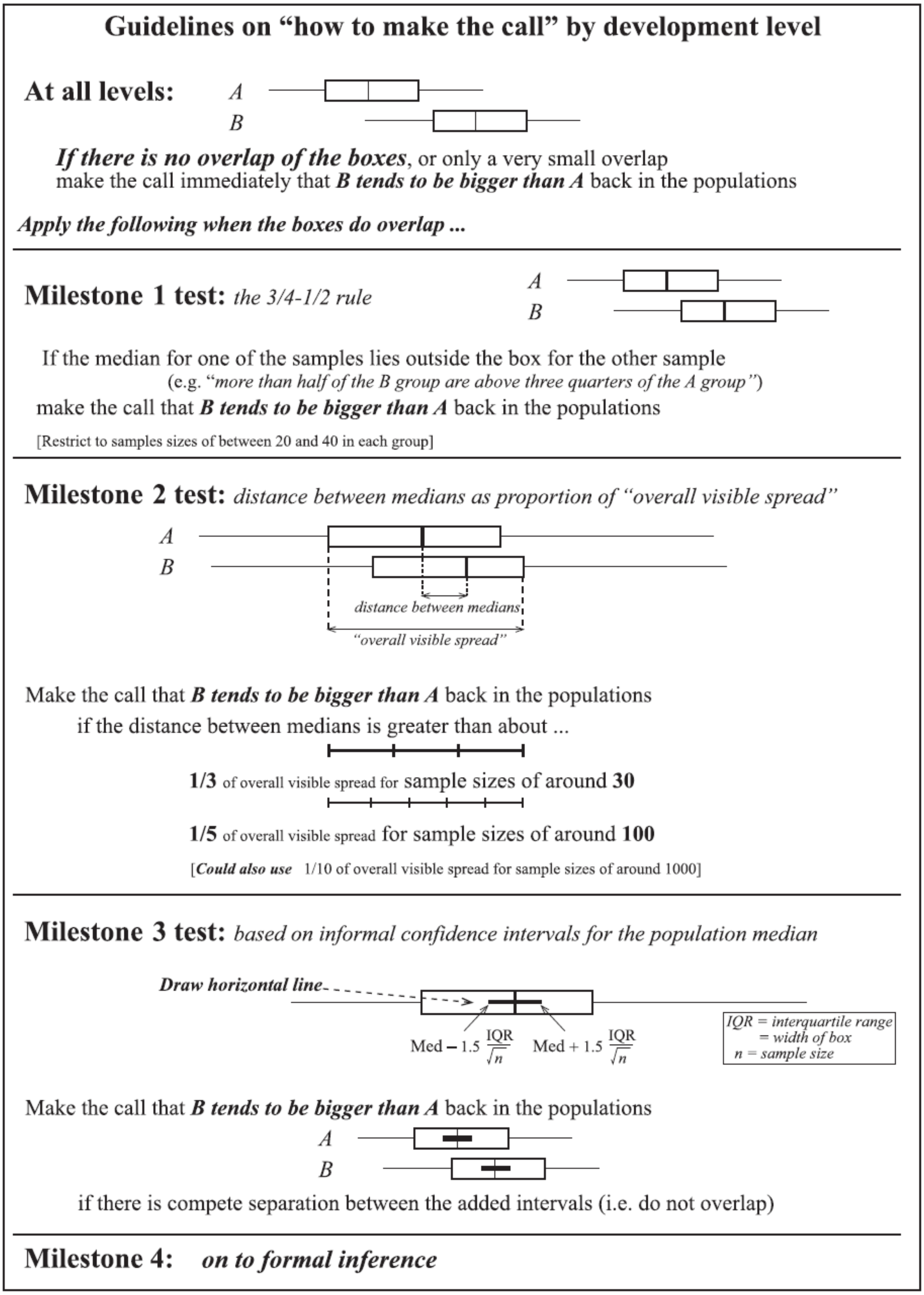

Figure 3. Robust Heuristics for Decision Makers at Different 'Milestones' of Statistical Sophistication (from Wild et al, 2011, in press)

More interesting, of course, is the likely size of the effect, and an understanding of Milestone 3 would be helpful. Wild et al offer the graphic shown on Figure 4 (their Figure 10, p15).

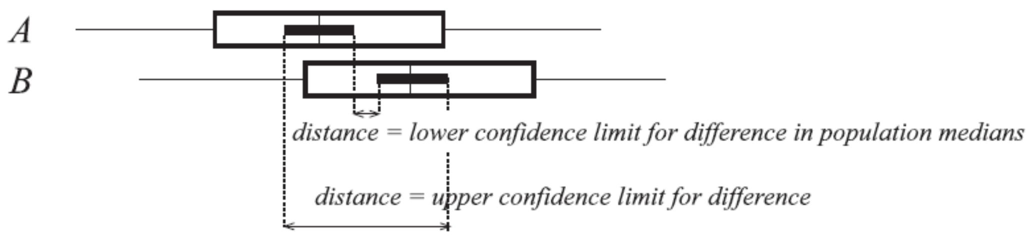

Figure 4: establishing Confidence Intervals for Differences between Population Medians (from Wild et al, 2011, in press)

On Nationally Mandated Crimes Against Statistical Inference

The measurement of school performance is a complex and difficult task, but one which should be addressed as part of efforts to improve the education of children. Here, we are concerned with just one aspect of this process – drawing justifiable conclusions from the evidence to hand. From the previous discussion, it is clear that the majority of primary schools are required to report evidence that no competent statistician would use to make judgements about their performance. Judgments about performance are being made on the basis of evidence that is unreliable. This 'evidence' has high stakes for schools; poor pupil performance is a limiting judgement in OFSTED Reports, so can downgrade a school's rating. At least as important, it can have direct and serious personal consequences for individuals. Frankel (2010) provides a number of examples of head teachers who have resigned from, or who have been forced to leave, their school on the basis of poor pupil attainment in a single year. The thrust of the article is to decry an accountability culture and 'football manager syndrome'. Our view is that accountability is a reasonable requirement to place on schools. However, accountability should be based on robust evidence and methods, not on methods that ignore basic principles of statistical inference.

Teachers in a very large number of primary schools are in the position of finding 'faces in the fire' – telling causal stories about events that have more to do with sampling error than with any aspect of pedagogy. Some teachers will be aware of the conceptual errors in the reporting system. Some, we suspect, will simply be aware that the reporting system does not represent school performance in ways they recognise, professionally. In either case, it makes little sense to create an assessment framework that involves a good deal of work, which invites conclusions that are unwarranted, and that is seen by the teaching profession as flawed.

We offer two suggestions. First, present performance data as boxplots, and promote the widespread use of the heuristics for statistical inference described by Wild and his colleagues. Second, focus the review of the performance of classes on the actual responses made by pupils – things they got right and things they got wrong. This sort of analysis can lead to fruitful professional discussion about classroom practice. External evaluation (such as OFSTED) could comment on the quality of the analyses, and the pedagogic content knowledge they evidence.

Acknowledgements: Thanks are due to Professor Chris Wild for allowing us to quote extensively from his work, and to James Nicholson for comments on an earlier draft of this paper. These acknowledgements do not imply that Chris and James are in complete agreement with what we have written.

References

Census at School New Zealand http://www.censusatschool.org.nz/

Census at School simulations http://www.censusatschool.org.nz/2009/informal-inference/WPRH/.

Department for Education (2010). *DfE: Schools, Pupils and their Characteristics: January 2010.* http://www.education.gov.uk/rsgateway/DB/SFR/s000925/

Department for Education (2010). *A School Report Card: Prospectus.* http://publications.education.gov.uk/eOrderingDownload/Report-Card_Prospectus.pdf

Frankel, H. (2010). *We need a result!* TES Magazine 22 October 2010. Available from http://www.tes.co.uk/article.aspx?storycode=6061359

Hattie, J. (2009). *Visible Learning.* Abingdon: Routledge.

League Tables http://www.education.gov.uk/performancetables/primary_09/pdf_09/840.pdf

Primary Schools Data Submission Guidance (2010). http://www.education.gov.uk/schools/adminandfinance/schooladmin/ims/datacollections/ks1softwareassessments/a0064510/key-stage-1-assessments-2010

National Strategies Guidance on Evaluation School Performance (2010) http://nationalstrategies.standards.dcsf.gov.uk/node/289919

RAISE on line https://www.raiseonline.org/maintenance.htm

Utts, J.M., and Heckard, R.F. (2007). *Mind on Statistics* (Third Edition). Belmont, CA: Brooks/Cole.

Wild, C. J. Pfannkuch, M., Regan, M. and Horton, N. J. (2010). Inferential reasoning: learning to “make a call” in theory. In *Proceedings of the 8th International Conference on Teaching Statistics* (ed. C. Reading). TheHague: International Statistical Institute. Downloaded from http://www.stat.auckland.ac.nz/~iase/publications/icots8/ICOTS8_8B1_WILD.pdf)

Wild, C.J., Pfannkuch, M., Regan, M., and Horton,N.J. (2011, in press). Towards more accessible conceptions of statistical inference. Journal of the Royal Statistical Society Series A 174(2), pp 1-23. Paper read before The Royal Statistical Society, downloaded from http://www.rss.org.uk/pdf/Wild_Oct._2010.pdf