REAL-WORLD CONTEXTS IN STATISTICS COMPONENTS OF UK MATHEMATICS EXAMINATIONS: AIMING FORWARD, WALKING BACKWARDS

James Nicholson and Jim Ridgway
University of Durham, UK
j.r.nicholson@durham.ac.uk

Reasoning with data is pervasive in society; evidence-based policy requires reasoning about complex data; many statistics curricula do not equip students to understand such reasoning. A new mathematics curriculum for upper secondary school (pre-university) students has been taught in England from September 2017, which includes assessment based on large-scale, pre-released data. Here, we critique the specimen examination papers, identifying inappropriate uses of statistics, and a failure to understand the rationale for working with large scale, authentic, data sets. We offer an account of how this poor state of affairs has come about.

INTRODUCTION

In England, the primary qualifications for college entrance are subject based examinations taken normally at age 18. Students will usually take examinations (called GCEs) in just 3 subjects (with a small number taking 4). At college level, many disciplines require quantitative reasoning. Over a decade ago, we argued that: reasoning with data is pervasive in society; evidence-based policy requires reasoning about complex data; and that the UK statistics curricula did not equip tomorrow's students to undertake such reasoning (Nicholson, Ridgway & McCusker, 2006; Ridgway, Nicholson & McCusker, 2007). The case is even stronger now, and major curriculum reforms (for teaching in 2017) offered a significant opportunity to align school experiences more closely with the developments in the world our students will inhabit.

The track record of using real data in assessing statistics within mathematics in the UK is extremely poor: analysis of the first module in probability and statistics from the four qualifications on offer in 2015 and 2016 shows 60 questions altogether (including probability based questions) with none quoting a source for data, and only two of the questions appear to be based on real observations.

As a result of the recent curriculum reform, all students in England will face examinations that are based on (almost) identical content; previously, there was choice amongst a range of modules. Now, all teachers will be teaching some material that is new to them, and in some cases the new material will represent a substantial proportion of the course. Ahead of the examination, students are required to work with 'a large data set' (LDS) in their course of study which will be used as the basis for at least some of the questions in the written examinations. Data sets are in the form of related tables, with multiple variables and many cases. Such data sets offer rich opportunities to explore real and important issues, and develop an understanding of how quantitative evidence can be used to try to answer real questions in authentic contexts. The curriculum documents articulate ambitions which are laudable: there was to be an increased emphasis on problem solving, on mathematical modelling and the use of technology was to underpin the course of study, along with the use of real data in statistics. Here, we address the extent to which these ambitions have been met.

WHAT MIGHT GOOD ASSESSMENT OF STATISTICAL KNOWLEDGE APPLIED TO REAL WORLD CONTEXTS LOOK LIKE?

Data have never been more easily available and technology has never been more affordable that at present. The trends in both areas are likely to continue, and should be reflected in curriculum changes. Here, we offer some guiding principles for good assessment:

- Assessment set in real world contexts should use authentic data
- The motivation for the analysis should be clear i.e. there should be a 'question of interest' to be explored (too often statistics assessment has been characterised by contrived contexts and artificially constructed datasets which allowed examiners to assess mastery of technique)
- Students should be assessed on their ability to use techniques appropriate for understanding multivariate data

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

- Students should have access to appropriate technology when answering questions
- Students should be asked to model non-linear relationships
- Students should be asked to articulate the meaning of their analyses
- Students should be asked to describe the practical consequences of their analyses
- If small samples are involved there should be a plausible reason as to why it was not viable to have a larger sample
- If a significance level other than 5% is to be used then there should be a plausible reason why it has been important to reduce the likelihood of Type I or of Type II errors in the context, (which candidates should be asked to explain).
- The assessment of probability should use contexts in which there is an obvious rationale for why a probability is to be calculated
- Any assumptions underpinning the use of probability distributions should be plausible in the context to which they are applied.

CAN THE NEW CURRICULUM DELIVER 'GOOD ASSESSMENT'?

The arguments used to justify the new curriculum are plausible. However, the techniques listed in the course specification do not facilitate the analysis of large multivariate data sets. Worse, some of the examination questions posed represent extremely poor statistical practice. In particular:

- There is nothing related to modelling relationships – other than interpreting straight lines of best fit (even calculations of straight lines using technology is excluded from the assessment). In the real world, non-linear relationships are much more common, and the use of *r*-squared as a measure of the proportion of the variance explained by a model could be usefully developed with widespread applications across other school subjects
- There is no encouragement to disaggregate the LDS by some characteristic (sex, geography, age) to enrich the analysis
- There is no encouragement to explore relationships between variables
- There is no attention to estimation – so nothing which would allow a meaningful exploration of the behaviour of the sampling distribution of summary statistics
- There is nothing about the mean and variance of probability distributions, or about what happens to the mean and variance of a distribution when a linear function of a random variable, or a linear combination of random variables, is taken – so again the sampling distribution of a statistic such as the sample mean will not be understood well enough to make the practical work of taking repeated samples a worthwhile educational activity. The graphs and summary statistics in the content specification only lend themselves to simplistic interpretations – single variable graphs, stretching to a comparative bar chart to include two variables, and bivariate graphs with straight line relationships
- No techniques are specified which allow any depth of nuanced interpretation (e.g. disaggregation)
- We have severe reservations about some of the questions which are posed in the Specimen Assessment Materials (SAMS) as we believe they represent extremely poor statistical practice, and impose an impossible burden if students are to prepare for the range of possible questions implied by the SAMS. Examples of this appear in the next section
- Statistical testing is required, even though there is nothing about estimation. There is no mention of Type I and Type II errors, so there is nothing to highlight the dangers of inference from small or from very large samples: sample sizes for tests were 7, 8, 9, 15, 18, 20, 25, 50, 50, 240, 450,
10 000, 12 144. To compound this poor statistical practice, four out of the thirteen contexts did not use random samples
- Lack of clarity: initially, the aim was that the LDS should be used to teach all the content, but the final criteria stopped short of articulating this explicitly. However, the specifications do recommend its widespread use, and do not identify topics where it should not be used
- Unwarranted procedures: the content document specified that students should 'be able to clean data'. All of the pre-release data is published by reputable statistical organisations such as the Office for National Statistics, and has been professionally cleaned before publication. Data

cleaning is much more nuanced and complicated than identifying data points and removing them simply because they are 'outliers', yet there is almost no other process which could be used to 'clean' one of these published data sets – students have no access to any of the other information that statisticians in these organisations will have used in the process of validating the data set before publication.

CRITIQUE OF SPECIMEN ASSESSMENT MATERIALS (SAMS)

Earlier, we identified some principles of what constitutes good assessment of statistics within a mathematics curriculum. We believe these principles are consistent with the recommendations of the GAISE report (2016) but that they articulate essential aspects of good assessment in a more concrete form. How do the SAMS match up to these principles?

We believe that there are serious shortcomings in the SAMS as exemplars for high stakes assessment instruments; we hope that many of the problems identified in this section will be avoided in live examinations. However, we believe that the curriculum content and structure cause some of the difficulties in producing good quality assessment. There are four specifications, from AQA, MEI, OCR and Pearson Edexcel. The specifications and SAMS are available for download from the websites of these bodies. See http://www.aqa.org.uk/subjects/mathematics/as-and-a-level for AQA, http://www.ocr.org.uk/qualifications/by-subject/mathematics/as-a-level-maths-from-2017/ for both OCR and MEI, and https://qualifications.pearson.com/en/qualifications/edexcel-a-levels/mathematics-2017.html for Pearson Edexcel.

We illustrate problems below; an analysis of all statistics questions in the SAMS can be found at https://community.dur.ac.uk/j.r.nicholson/ICOTS10.htm

- MEI AS paper 2 Q9 uses data from their LDS which consists of data about countries from the CIA World Factbook with the addition of data about medals won in the London 2012 Olympics. Part of the question and mark scheme is shown in figures 1 and 2 below.

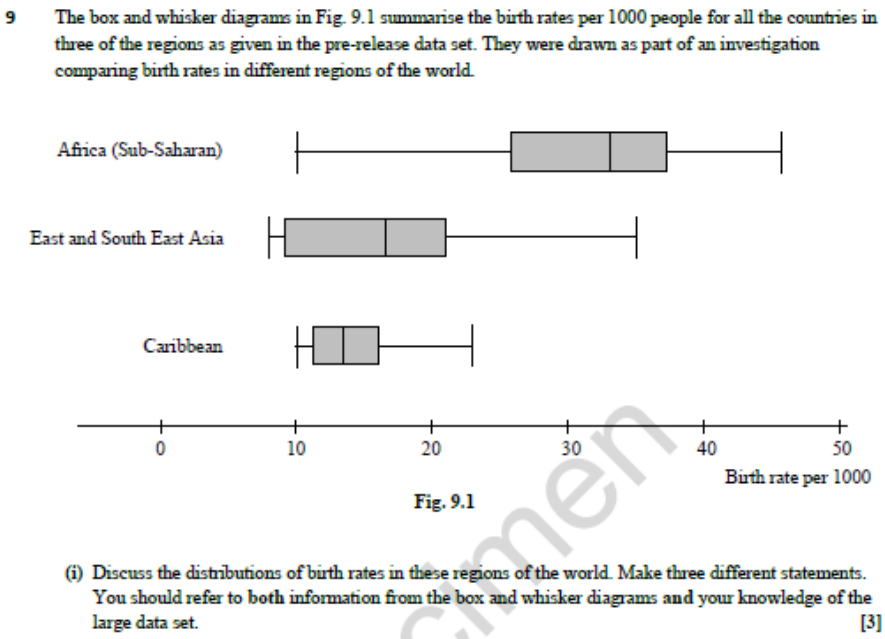

Figure 1: MEI AS paper 2 Question 9 a)

9	(i)	E.g. There is a greater spread of birth rates for countries in sub-Saharan African than for countries in the Caribbean E.g. The range for countries in Africa is greater than for countries in East and South East Asia but this could be caused by outliers as the IQRs are similar E.g. sub-Saharan Africa has a mixture of economically rich and poor countries resulting in a large IQR E.g. Countries in East and South East Asia tend to have higher life expectancy than countries in sub-Saharan Africa so their populations are older, on average, and have lower birth rates	B1, B1, B1 [3]	2.2b 2.2b 2.2b	B1 Correct relevant comment that can be inferred from the source material B1 Distinct correct relevant comment that can be inferred from the source material B1 Third distinct relevant comment that can be inferred from the source material (this mark is only available if the candidate's comments include reference to both features of the LDS and fig 9.1)

Figure 2: MEI AS paper 2 mark scheme for 9 i)

Figure 1 shows country birth rate data for 3 sub-regions and candidates are asked to make causal statements using information not in the given diagrams. The scale of what you might be asked here is unreasonable: there are 14 sub-regions, and 9 variables from the CIA World Factbook. Moreover, the reasoning about causality given in the mark scheme (figure 2) is deeply flawed. The second response focuses on the spreads in comparing Africa and Asia which is a much less important feature than where the majority of birth rates lie in the two regions, and the comment about outliers is much less important than the dramatic difference in skewness of the two groups. In the third and fourth responses, we don't think these are reasonable causal assertions to make from the evidence. All the specifications state that candidates need to know that correlation does not imply causation, yet they are repeatedly required to make causal statements from associations. Further, there is an issue about how many 'facts' candidates are supposed to know, in order to make assertions like this from the LDS. Part iii) asked candidates to describe correlation in a scatter diagram for the full LDS, where the relationship was obviously non-linear.

- AQA AS paper 2 Q14 raises similar concerns over the scale of demand placed on candidates as to their recall of factual knowledge in the LDS. It is a multiple choice question to identify a region from a stacked bar chart of sales of butter and margarine for 2014. The LDS has 14 years' data on 110 products, and each region is contained in a separate sheet, with 10 sheets altogether. To compile this diagram would take a modest amount of time in extracting the data, but the number of equivalent possible questions which could be asked is not reasonable, and it rewards factual recall – not credit for statistically-based reasoning.
- The context for MEI AS paper 2 Q10 was a journalist's feeling that the proportion of on-time train arrivals was overstated. Train companies are required to publish all the data on arrival times. The question asks whether a random sample of 18 trains has a lower proportion of on-time arrivals. Students are asked to use a 1% probability level - why 1%? Why is a small sample used? Indeed, why is a sample used at all when population summary statistics are available?
- OCR AS Paper 2 Q12 asks for a 10% one-tail test of the proportion of 450 patients suffering side effects (what is the rationale for a 10% one-tail test?). Having performed the test, candidates are informed all the patients came from the same hospital and asked to comment on the validity of the model used. This is questionable practice – candidates are told to perform an inappropriate procedure.
- OCR AS Paper 2 Q10 tells candidates to use a $B\left(7, \frac{3}{8}\right)$ distribution to model the number of appointments starting late on one day for a single consultant. It is hard to think of a context in which assuming independence of trials would be less reasonable – this is extremely poor practice in a qualification which has a particular emphasis on modelling.
- Pearson Edexcel AS paper 2, Q4 uses an LDS on weather, with Sara taking a random sample of 11 days from one month at one location from the whole data set (so sampling was restricted to choosing 11 from 31 cases in an LDS with 2944 cases), and asked to show that one rainfall reading is an outlier. Candidates are asked to give one reason why she might include, and one why she might exclude. this reading. A scatter diagram with the reading excluded is given, along with the equation of the regression line, and candidates are asked to give an interpretation of the correlation, and of the gradient of the regression line. An earlier version of this question was published which

asked candidates to use the regression equation to calculate a value Sara could use to replace the outlier. While this was removed in the final SAMS published, teachers were not alerted to the fact that this is extremely poor statistical practice.

- Pearson Edexcel A level paper 3 Q2, talks about 'a random sample of 9 consecutive days' from a UK town in July and asks candidates to perform an hypothesis test on the correlation coefficient between the daily mean windspeed and the daily mean temperature. This is nonsensical: 9 consecutive days cannot conceivably be described as a random sample, so conducting a hypothesis test is inappropriate. A 'randomly selected set of 9 consecutive days' would be the accurate description of the sample, and if this language had been used then the examiners might have recognised that they could not set a hypothesis test based on it.
- MEI A level paper 2 Q16 part iii) asks: *Decide whether the maximum increase in life expectancy from 1974 to 2014 is an outlier. Justify your answer.* The information provided is in a boxplot which has no scale shown on it, and a table of the values in the boxplot - but what is the purpose of this question? So the examiner can tick off another technique that has been assessed? Candidates do nothing with the outcome, so why determine whether or not it is an outlier (it is an outlier – just)? In the same question, part iv) asks candidates to estimate change in life expectancy for two countries, having told candidates only what interval the two countries are in for one of the time periods given. Since the actual values are known, under what circumstances would this be a useful technique or skill to have? (further, requiring an answer to 1 decimal when one value is only known to lie in an interval of width 5 is simply bad practice). Figure 3 shows the scattergraph for parts v) and vi) – both axes represent life expectancy but the axes do not have equal aspect. The graphic clearly shows that the homoscedasticity assumption is not close to reasonable, yet candidates are required in part v) to use the line of regression to estimate a life expectancy in 2014. Part vi) asks how many countries had a drop in life expectancy from 1974 to 2014. This requires the use of the line $y = x$, which normally goes through the origin at an angle of 45^0 - here it does not go through the bottom left of the graph and it makes an angle of 31.6^0 with the horizontal (we have shown the line $y = x$ as a red dotted line on the graphic). The commentary provided when this question was in an earlier draft form says 'pupils have to decide how to solve the problem' but this part question represents a very impoverished view of 'statistical problem solving' in our view. Indeed, if we wanted to 'solve this problem' we would look in a table to count the number of negative values – not do it graphically where to get a reliably precise answer involves carefully constructing a very unusual line, with counter-intuitive properties, and then counting symbols.

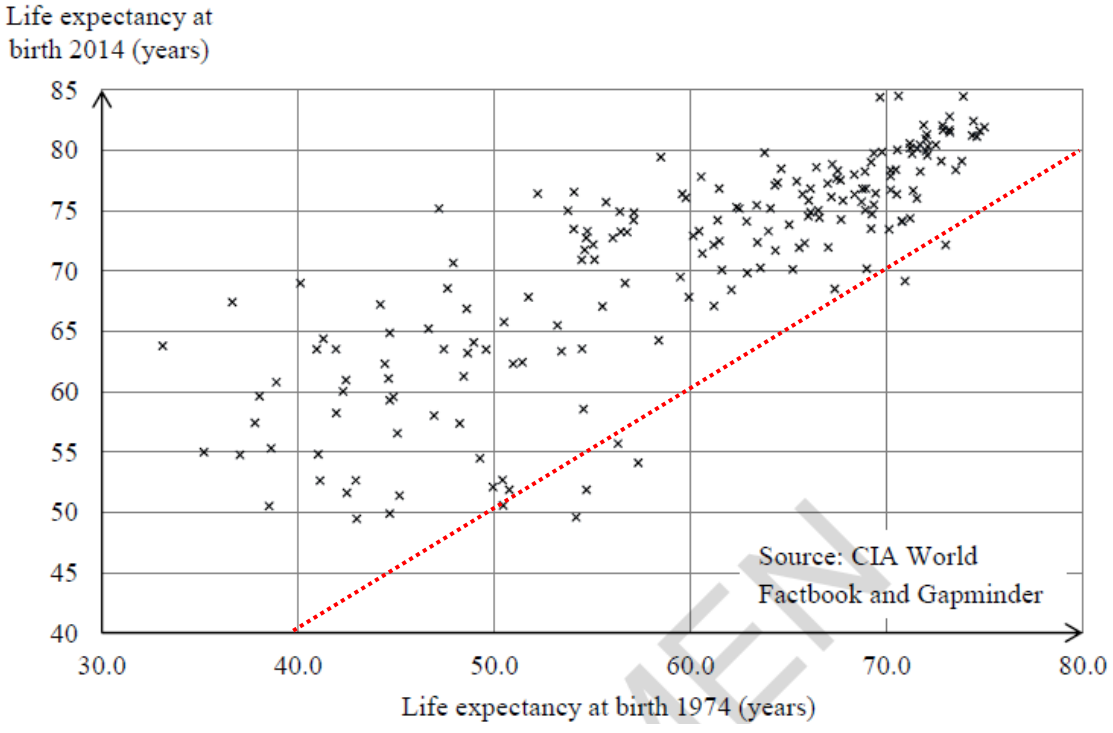

Figure 3: MEI A level paper 2 data used in question 16 v) and vi) showing the line $y = x$

DISCUSSION

The SAMS fail to capitalise on the potential benefits of using large scale data sets, and reflect serious statistical misconceptions. How has this state of affairs come about? The new qualifications are the product of two separate regulatory processes, overseen by different government departments. First the Department for Education (DfE) produced the curriculum document that set out the content (https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/516949/GCE_AS_and_A_level_subject_content_for_mathematics_with_appendices.pdf) and imposed the use of the LDS. Awarding organisations then submitted proposals for qualifications to Ofqual for accreditation, including how assessment would be carried out, together with SAMS. The DfE document is badly flawed: the changes introduced in the statistics section of the new GCE mathematics qualifications are well-intentioned – one might think it is trivially true to say that real data should be asked to answer real questions, but on the evidence of the previous curriculum and indeed the new curriculum, it is not. Previous examinations used almost no real data sets, even when there was freedom to do so. So there was very little experience of asking real questions about real data on which to build. An appropriate first step would have been to require new examinations to use real data.

The introduction of a compulsory pre-release LDS was a radical innovation. To our knowledge, it has never been used before, anywhere in the world. There were other important changes to the qualification as well, and before submissions were made to the accreditation process, Ofqual set up the A level Mathematics Working Group to produce a report on mathematical problem solving, modelling and the use of the LDS. In the Terms of Reference, the group was tasked with producing a range of exemplar questions and marking principles, with supporting commentary, stating: *These questions will focus primarily on mathematical problem solving and will cover the mathematical modelling process and statistics in relation to large data sets* (see https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/481857/a-level-mathematics-working-group-report.pdf). The working group report contains 6 exemplar questions on problem solving, and 4 on modelling, all complete with mark schemes and commentary, but did not provide a single, sample, LDS with associated questions as proof-of-concept to illustrate how good statistics assessment materials can be based on an LDS. Nicholson (private communication, 2016) wrote to the DfE arguing that, with the techniques specified in the content, he believed it would not be possible to set worthwhile assessment items based on an LDS which would reliably and validly reward time spent working on that particular data set. He expressed the view that it was inadvisable, *at best*, to proceed with such a radical and completely untested innovation when the advocates had not provided a proof-of-concept example. The above analysis of the SAMS suggests that these concerns were well-founded.

The short timescale of implementation meant that both the draft specifications and draft SAMS were placed in the public domain at the same time they were submitted for accreditation. Changes were made to draft SAMS when resubmissions were made. When changes were made, no commentary was given about the underlying reasons - so teachers were given no information about why a change had been made - did it reflect bad statistical practice, or was it changed for technical reasons such as coverage of topics, balance or level of demand etc.? We believe that this is wrong in principle, and harmful in practice: the development and accreditation of specifications should be carried out first, and only the accredited materials should appear in the public domain. This process should be completed at least a year before first teaching so that teachers can prepare for changes, and publishers have a realistic time frame in which to produce teaching resources for the new specifications. What has happened in this instance provides a salutary lesson as to why this is important. If current practice is to continue, there is an argument that, on the grounds of transparency, Ofqual should publish its reasons for rejecting a submission, including where questions were flawed.

These specifications are accredited for an unspecified period: realistically they are likely to be in place for a minimum of 5 years. We believe that work needs to start in the very near future on revisions to be introduced for first teaching in September 2022, which should be put through the

full accreditation process by September 2021. We believe that the use of real data should be required in assessment, including making comparisons between sub-groups of a larger population. However, we believe that the use of pre-released multivariate data is not appropriate because the obstacles to valid and reliable assessment using it are insurmountable without a radical change to curriculum content, and to the barriers associated with providing appropriate technology to support high-stakes assessment. We believe the current SAMS include many examples of extremely poor assessment items: we hope that, in the short term, questions will be subject to expert critique before release. In the longer term, we hope to see curriculum changes that require appropriate statistical techniques to be applied to authentic data sets, and assessment systems that reward these statistical skills.

REFERENCES

Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., ... & Wood, B. (2016). Guidelines for assessment and instruction in statistics education (GAISE) college report 2016. *Alexandria, VA: American Statistical Association.[Online: www. amstat. org/education/gaise]*.

Nicholson, J.R., Ridgway, J. & McCusker, S. (2006). Reasoning with data – time for a rethink? *Teaching Statistics, 28*(1), 2-9.

Ridgway, J., Nicholson, J.R., & McCusker, S. (2007) Teaching Statistics – Despite its Applications. *Teaching Statistics, 29*(2), 44-48.