Problematising high-stakes assessment in statistics

La problematización de la evaluación de alto nivel en estadística

Jim Ridgway and James Nicholson

School of Education, University of Durham, UK

Abstract

Statistics emerged as a discipline to address pressing practical problems. In the UK, this has not been reflected in school statistics curricula, where students often work with small-scale invented data to develop mastery of statistical technique. Recent curriculum reforms set out to improve this situation; students are expected to work in class with a large authentic data set, and to demonstrate appropriate skills on high-stakes assessment. Here, we analyse all the first set of examination papers containing statistics for the new GCE qualification, and also questions using statistical graphs from the GCSE qualifications in summer 2017. We show that there is very little emphasis on statistical skills such as interpreting data and drawing conclusions, and a great deal of emphasis on technical skills. Contexts are (for the most part) banal. Several questions ask students to use inappropriate procedures. We believe systemic flaws have resulted in assessment which is not fit for purpose. We call for curriculum reform, and offer examples of how things might be done better both in curriculum and in assessment.

Keywords: high-stakes assessment, curriculum reform, large-scale data sets, errors, misconceptions

Resumen

La estadística surgió como una disciplina para abordar problemas prácticos apremiantes. En el Reino Unido, esto no se ha reflejado en el estudio de la estadística en la escuela, donde los estudiantes a menudo trabajan con datos a pequeña escala inventados para desarrollar el dominio de las técnicas estadísticas. Las reformas curriculares actuales se proponen mejorar esta situación; se espera que los estudiantes trabajen en clase con un gran conjunto de datos auténticos, y que demuestren habilidades apropiadas en la evaluación de alto nivel. En este trabajo analizamos el primer conjunto de documentos de examen que sobre estadística para la nueva calificación GCE, y también las preguntas que utilizaron gráficos estadísticos de las calificaciones GCSE en el verano de 2017. Mostramos que hay muy poco énfasis en habilidades estadísticas, como la interpretación de datos y la extracción de conclusiones, y un gran énfasis en las habilidades técnicas. Los contextos son (en su mayor parte) banales. Varias preguntas piden a los estudiantes que usen procedimientos inapropiados. Creemos que estas deficiencias sistémicas han resultado en una evaluación que no es apta para el propósito educativo pretendido. Hacemos una llamada a la reforma del currículo, y ofrecemos ejemplos de cómo se podrían hacer mejor las cosas tanto en el currículo como en la evaluación.

Palabras clave: evaluación de alto nivel, reforma curricular, grandes conjuntos de datos, errores, conceptos erróneos

1. Introduction

High-stakes assessments are assessments which facilitate or block student access to careers or to future study pathways. In England, students take high-stakes examinations (GCSEs) at the end of compulsory schooling (at age 16 years) in a number of academic disciplines such as English, Geography and Mathematics, and if they want to study at university take further (discipline-based) examinations (GCEs) at age 18 years. There are syllabus specifications for every course, but a major source of information about exactly what is to be learned is the examination papers that are set then released into the public domain, once the examination cycle in any year has been completed.

Teaching statistics in school is problematic for a number of reasons. One is the nature of the discipline itself. Statistics emerged primarily as a practical subject; people with

Ridgway, J., and Nicholson, J. (2019). Problematising high-stakes assessment in statistics. In J. M. Contreras, M. M. Gea, M. M. López-Martín y E. Molina-Portillo (Eds.), *Actas del Tercer Congreso International Virtual de Educación Estadística*. Disponible en www.ugr.es/local/fqm126/civeest.html

diverse expertise and backgrounds came together to collaborate in order to solve real problems, often by inventing new mathematics. Pullinger (2014), for example, points to the diversity of expertise amongst the founders of the Royal Statistical Society. This gives rise to the problem of where statistics should appear in the curriculum – it is central to social and applied sciences, but does not have a natural home in any single discipline. The application of statistics requires some technical fluency – again, this mitigates against a natural home in any social science discipline, where teachers' fluency in mathematics cannot be depended upon. Mathematics teachers can be expected to be able to acquire the technical skills necessary to use statistical technique appropriately, but may have neither the skills nor the motivation to relate mathematics to real-world contexts. The lack of embedded subject contexts for statistics within mathematics has resulted in teaching and assessment which has often been based on artificial (toy) data sets (e.g. Cobb, 2015). This is not necessary; we will provide examples of realistic contexts addressing substantive questions of interest to show that it is possible to improve on the current classroom and assessment practices.

A second problem is that the use of technology is essential to the practice of statistics. Teaching statistics is made much easier when technology is used. One might hope that in any assessment of attainment in statistics, students would have access to appropriate technology. However, for a variety of practical reasons, statistics in England is assessed primarily via paper-based tests with access only to a scientific calculator.

A related third problem is that curriculum specifications often hark back to the days when calculations had to be done by hand, with the result that students are required to learn techniques that are always automated in professional work.

A related fourth problem is the speed with which the digital world is developing. New ways to display data continue to emerge – (see Ridgway, Nicholson, Campos, & Teixeira (2018) for a review). New sorts of data are available – such as images, texts, twitter streams, and transactional data from fitness monitors or consumer purchases. Of direct relevance to the school curriculum is the ready availability of large-scale authentic data sets, such as data on the UN Human Development Indices, Census data, and the like. There is a willingness to use such resources, but some barriers inhibit their use.

A fifth major problem is the statistical expertise of people writing examination papers, and those involved in the quality assurance associated with high-stakes examinations.

In this paper, we focus primarily on statistical graphs, both in examination questions on high-stakes tests and on how the curriculum might develop to provide more appropriate skills for our young people; we believe that the issues we address in relation to these visual displays have parallels in those other areas.

2. Examining examinations

In England, examination papers are created by three independent Awarding Organisations (AOs). We analysed the GCSE papers set by each of these AOs in June 2017. In the 18 papers (3 papers at two tiers for 3 AOs) there are 20 distinct questions using statistical graphs (4 of these appear twice - being used at both Foundation and Higher Tier by one AO). Our analysis (see Table 1) showed that there are only four questions where any sort of inferential reasoning is used (even when we take a very broad interpretation of inferential reasoning) with a total of 6 marks out of 81 on the parts using inferential reasoning. Further, many of the questions require candidates to use inappropriate techniques, or to engage in activities which have no sensible purpose.

Table 1. Content of examinations

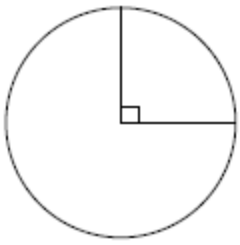
Type of work assessed	Number of marks	Percentage of marks
Construction of graphs	19	23%
Basic reading of graphs	9	11%
Calculation using information from graph	31	38%
Identifying errors in a given graph	5	6%
Identifying outliers	2	2%
Interpretation (describing correlation or comparing distributions)	9	11%
Any sort of inferential reasoning	6	7%
Total	81	98% (due to rounding)

Most of the marks are allocated for the deployment of mathematical techniques such as simple calculations and constructing graphs. A variety of tasks are set in realistic contexts, but these contexts are almost all dull and contrived, and students are rarely asked to draw conclusions that have any interest to anyone, for example: interpreting pie charts and tables of goals scored in football; calculating the interquartile range of hours spent on homework; analysing voting by sex; calculating cut-offs for examination marks and customer spending in a shop; critiquing a display about CD sales; looking for errors in a plot of ice cream sales against temperature over 5 days – by a hypothetical ice cream seller.

3. Illustrative problems with high-stakes assessment items

Example 1: OCR summer 2017 Paper 1F Q9 - An ill-conceived task requiring puzzle solving

We are told that Jorge recorded the scorers of 120 goals and started to draw a pie chart to show the results. Figure 1 shows the diagram given, along with the table giving information about the other players who scored goals. Candidates were asked how many goals Simon scored and are then instructed to complete the table and the pie chart.



Name of scorer	Number of goals	Angle of sector
Wayne	48	144^0
Harry	5	
Obi		72^0
Antony		

Figure 1. Information given in *OCR summer 2017 Paper 1F Q9*

Pie charts use multiplicative reasoning, which is a key aspect of the GCSE curriculum. However, here candidates are being asked to do things here which have little to do with representing or interpreting data. Candidates were asked to complete a table and a pie chart with information that is presented in ways that are never used outside the examination room. The first sector is shown on the pie chart as a right angle, and students are told that there 120 goals scored altogether; they are then asked to calculate the number of goals Simon scored. Then a table is given of other people's scorers (excluding Simon), and candidates are asked to: work out sector angles using the number of goals; work out the number of goals from sector angles; then (for the last row of the table) work out the

number of goals Antony scored, using their earlier calculations. Adding up the angles in the table gives 270 degrees (not 360) because Simon is not in the table.

This question has nothing to do with statistics. A table of the number of goals scored is all that is required to describe the situation of interest; a pie chart would add nothing of value. The question requires the application of some logic to some numbers, and nothing more. Worse, it communicates the message that statistics itself is of no practical value - in what circumstances would anyone ever be faced with the sorts of calculations and reasoning required here?

Example 2: OCR summer 2017 Paper 3H Q10 – An invalid activity

Candidates are given a table of the amounts of money spent in Ana's shop; intervals are unequal. They are asked to construct a histogram, and then to estimate a cut-point to identify customers in approximately the top 25% of spenders. These customers will then get a money-saving voucher. The table of spending patterns would have been derived from raw data (obviously); the activity of using aggregated data to do a sum best done on disaggregated data is (at best) fatuous. If only the grouped frequency data were available then, graphically, a cumulative frequency curve would be a much more appropriate approach than using linear interpolation from a histogram. The question rewards students for applying the wrong technique to a supposedly-realistic context.

If candidates use interpolation to calculate the cut-off value (the top 25% of customer spends) the result is £17.22. The mark scheme allows anything from £17 to £18 inclusive with 'valid working and justification'. We note that there is a missed opportunity to emphasise that, in this context, the suggested value should be a rounded value – here, even a candidate giving an answer of £17.22 would be awarded full marks.

Example 3: Edexcel AS mathematics summer 2018 paper 2, Q4 - Poor implementation of good intentions

An important and potentially valuable development in the curriculum is the idea that students should be required to work in class with large authentic data sets (LDS). However the current implementation of this has a number of very significant systemic flaws which have resulted in assessment which is not fit for purpose in high stakes examinations. Almost all students and their teachers have had little or no experience of working with real data, not even in the context of examination questions. The only statistical techniques in the course are: univariate descriptive statistics; bivariate data associated with straight lines; sampling and cleaning data; working with Binomial and Normal distributions; hypothesis testing for Binomial probability; mean and variance of the Normal; and correlation coefficients. Students and teachers are now required to work with a pre-released LDS which has multivariate data in complex contexts. Each of the data sets used by the different AOs contains variables of types which have not been a part of the prior curriculum and which should not be analysed using the techniques currently in the content specification. For example, maximum daily wind speed, death rates for countries with wildly different population sizes, and average consumption per person per week by region.

We believe that the national curriculum needs a radical reshaping so that students work with real data at primary school and develop confidence and fluency is applying standard basic statistical techniques to scale variables, supported by the use of technology. The emphasis should be on choosing appropriate displays and on describing stories in the display, including critical evaluation of whether data in a display is consistent with a given proposition. Currently, textbooks and assessments almost never address issues

concerning how data is best represented; yet when technology allows the process of construction to be automated, this is the skill that students have most need of as they move into professional life. In Appendix 1 we offer an example of some curriculum materials to illustrate what we mean about teaching students appropriate criteria for deciding the most appropriate graphical display.

The Edexcel AS mathematics summer 2018 paper 2, Q4 uses data from a pre-release large data set containing meteorological data for a number of locations for six months in each of 1987 and 2015. The examination is supposed to reward candidates who have spent time working with this data set.

The means and standard deviations of the daily mean wind speed for five months in 1987 for one location from the large data set are given in Table 2 below. The data were not in month order. A diagram showing boxplots of the data for the same five months was provided, and is reproduced below in Figure 2. Candidates were asked to suggest, giving reasons, which of the months in Table 2 is "most likely to summarised in the box plot marked *Y*."

Table 2. Summary statistics for daily mean wind speeds

Month	A	B	C	D	E
Mean	7.58	8.26	8.57	8.57	11.57
Standard Deviation	2.93	3.89	3.46	3.87	4.64

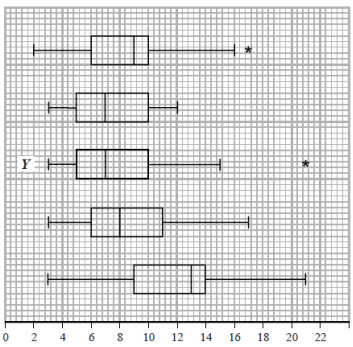


Figure 2. Boxplots for data in table 2

The activity is inherently silly – students are asked to map a 2 parameter summary (mean and variance) of an unknown distribution onto a 5 parameter description (boxplots). The shape of the distribution is not known to the students (we have taken the distribution of windspeeds form the LDS and present them in Figure 3).

The (provisional) mark scheme reads:

> Y has low median so expect lowish mean (but outlier so > 7) and Y has big range/IQR or spread so expect larger st.dev Suggests B.

The notes to the mark scheme say:

> M1 [one mark] for a comment relating to location that mentions both median and mean and a comment relating to spread that mentions both range/IQR and standard deviation and leads to choosing B, C or D

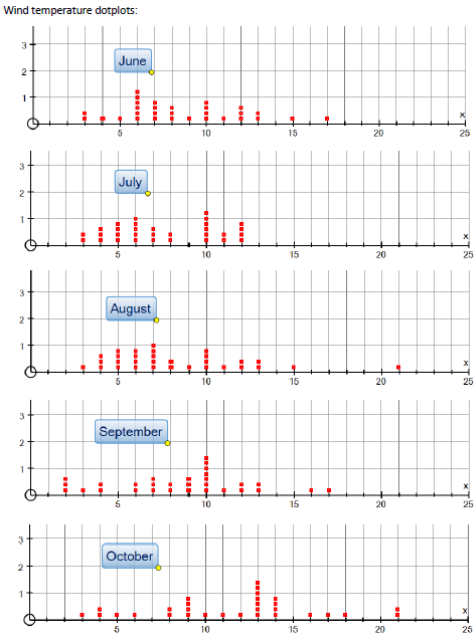


Figure 3. Daily mean wind speeds for the months summarised in table 2

With access to the raw data it is a relatively simple matter to explore what the upper and lower limits of the mean are for any boxplot – by moving values not affecting the quartile or outlier calculations as far as possible down and up. Table 3 shows the results of this exploration – in the same order as displayed on the boxplot:

Table 3. Possible values of the mean consistent with the boxplots shown in figure 3

		Actual	Min	Max
D	Sept	8.57 (3.87)	7.60 (3.88)	9.77 (4.17)
A	July	7.58 (2.93)	6.52 (2.79)	8.26 (2.89)
B	August	8.26 (3.89)	7.19 (3.85)	9.16 (4.29)
C	June	8.57 (3.46)	7.43 (3.43)	9.90 (4.35)
E	Oct	11.57 (4.64)	10.23 (4.75)	13.83 (4.72)

The examiners' report on this question said:

> Part (ii) was very challenging and required some careful thought and inference. Some confused mean with median and assumed that C and D were the 2nd and 3rd box plots in the list since they had the same medians. The intention was that they should identify E as the 5th box plot and A as the 2nd (it has a low mean and standard deviation and clearly has the smallest range on the box plots). Looking at the other 3 they should notice that the range is large and the outlier too would suggest a large standard deviation. The box plot Y has the smallest median of these 3 so a lower mean is suggested which points to B. We were looking for some explanation using correct terminology so we expected references to the mean and median and the standard deviation and a suitable measure of spread from the box plots.

This sort of reasoning is deeply problematic, and depends upon strong assumptions that may well be wrong. Anscombe's famous data quartet highlights the dangers of making assumptions about what the data looks like based solely on summary statistics. He created 4 strikingly different data sets (see Figure 4) where x and y have the same mean and variance, and the same regression line and correlation. The scattergraphs show how misleading summary statistics can be about distributions and relationships between variables.

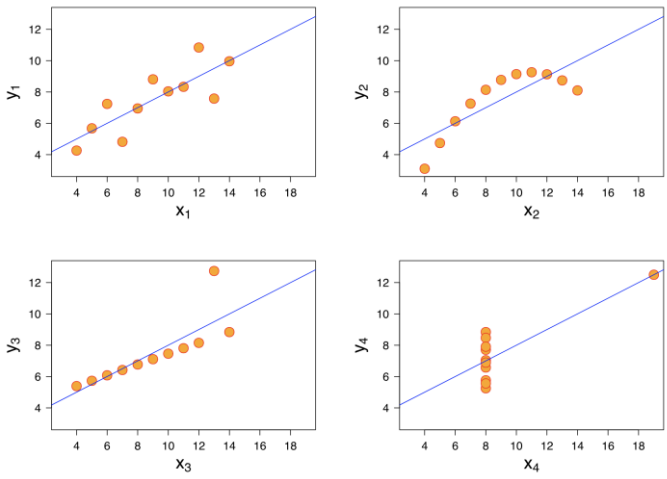


Figure 4. Data from Anscombe (1973)

It is plausible to argue that the bottom boxplot in figure 2 would be E with the high mean, and just about plausible to argue that the second boxplot would be A with the low mean. The possible values for the mean for the 3 months with the quartiles defined by the three other boxplots are in the intervals (7.60, 9.77), (7.19, 9.16), and (7.43, 9.90) and the standard deviations are quite similar. The means of the three months not deemed unlikely to be Y are 8.26, 8.57 and 8.57: all of which are well within the intervals in which the mean for each of the three months can be deduced from knowing only the quartiles and any outliers. Even if the distributions within any given month were consistently well behaved it would be assessing factual recall (rather than any aspect of the course) that candidates would be being rewarded for to provide the reasoning indicated in the mark scheme. Since the dotplots shown in Figure 3 quite clearly indicate that the 30 or 31 values in a month are haphazard, a candidate who has actually explored this facet of the LDS would be penalised rather than rewarded for their time spent on the LDS.

It appears from the mark scheme that the examiners would accept any ‘sensible’ explanation. However, the question reads as if the student should be able to identify the ‘correct’ answer from the information given. Conscientious candidates who feel they know this area of the course may spend a considerable time trying to work out how to identify the ‘correct answer’.

This question provides another example where students are asked a question that no data analyst would ask – namely to draw conclusions from summary data, where a (very simple) analysis of raw data would give the answer. Students are also required to make assumptions that may not be justified. Such assessment items are damaging in a number of ways. Students are being asked to work with data in inappropriate ways, in order to demonstrate some (faulty) understanding of the relationships between different summaries of data.

Teachers are put in a difficult situation – if they are statistically sophisticated, they will be dismayed – and perhaps demoralised by having to prepare students to engage in ‘trick behaviour’. If they are statistically naïve, or unconfident in the knowledge of statistics, they may acquire some statistical misconceptions, and are unlikely to develop any greater confidence in their statistical knowledge.

Apart from the substantive issues raised about this question, an important concern is that the curriculum specification makes no reference to the mapping the relationship between summary statistics.

4. When 'experts' get the statistics wrong – in a very public way

It is reasonable to expect that documents about statistics in schools from authoritative sources should not contain major errors. Here we show some problems contained in a document entitled *The Future of Statistics in our Schools and Colleges* (Porkess, 2012).

Figure 5 below shows the entries in GCSE Statistics examinations in the UK between 2003 and 2011. There is a strong pattern in the entries – they rise steadily from 2003 to a peak in 2008 and then the pattern abruptly reverses with a steady decrease in the next three years.

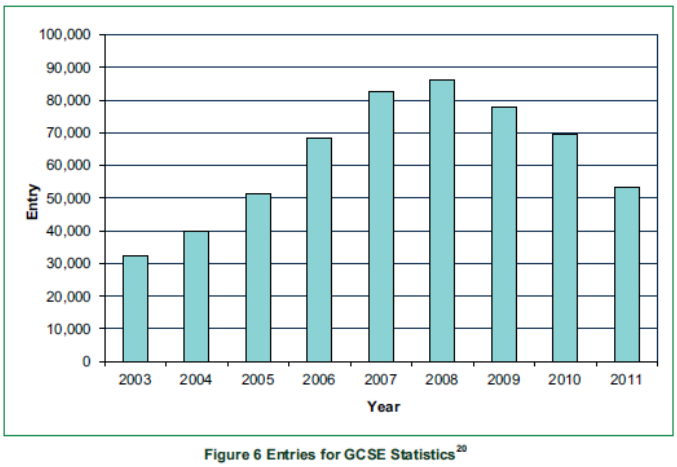


Figure 5. Taken from Porkess (2012, page 29)

However, the commentary in the report says:

> This is supported by Figure 6, which shows the uptake of GCSE Statistics in recent years; the very large variation suggests that factors other than an appreciation of statistics are involved.

Indeed there is a large variation in the entries; if the time series had looked like the graph shown in Figure 6 below (the same values but in haphazard order) then this comment would be entirely appropriate – the mean entry was around 62,000 with a standard deviation of around 18,000 for the years 2003 – 2011. However, ignoring the strong pattern in time series data is serious mistake.

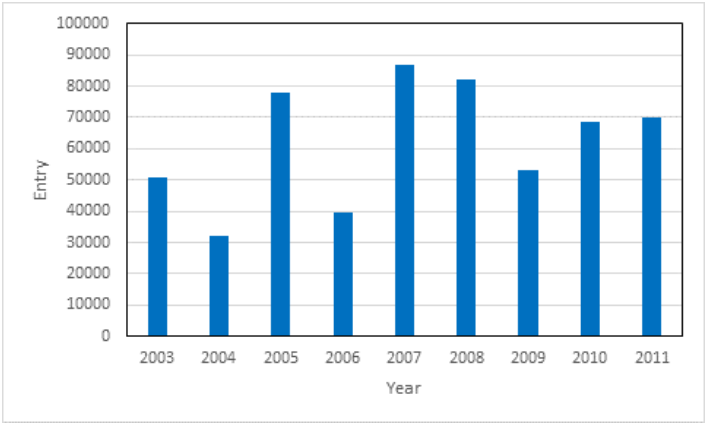


Figure 6. Data from Porkess (2012) re-ordered

Coursework was a compulsory part of both GCSE Statistics and GCSE Mathematics between 2003 and 2008, and the coursework used in GCSE Statistics could be used as one of the two pieces of coursework needed for GCSE Mathematics. In 2009 (and

following years) coursework was removed from GCSE Mathematics but stayed in GCSE Statistics. This offers a plausible (but not conclusive) explanation for the changes in GCSE entries. The major point being made here, though, is that authoritative documents about statistics education should show (teachers and others) good practice in working with graphs.

5. Addressing the problems

Some of the problems we have identified here relate to a lack of expertise in setting appropriate examination questions. Some of the questions which have been set reveal a lack of statistical understanding; some ask students to engage in procedures that no one working with data would ever use; some appear to have been set simply to fit with ill-thought-through curriculum specifications. Addressing the problem of poor examination questions is straightforward in principle. More knowledgeable examiners, and better quality assurance processes are required. The issue of an ill-specified curriculum is harder to address, and requires concerted effort with government departments to achieve desirable goals.

Some changes are urgent we believe. The current emphasis on constructing graphs with nothing about the relative strengths and weakness of different representations for different purposes is inappropriate. In Appendix 1 we offer an example of what curriculum materials could look like to address this.

6. Towards better assessment and better curricula

Using better examination questions

Analyses of questions on high-stakes examinations revealed little use of authentic data, and no attempt to use data for practical purposes. Here we offer some examples about how this might be done, with a commentary following each question about why it differs from current items.

> Q1: Walker's disease is a rare tropical disease, known to be present in only 0.1% of the population. A new screening test has been analysed, and there is a 98% probability of testing positive when the person tested has the disease, and only a 0.2% probability of testing positive when the person does not have the disease.
>
> A person is selected at random from the population and given the screening test.
>
> a) What is the probability that the person will test positive?
>
> b) What is the probability that the person does not have the disease, given that they test positive?
>
> c) Jane is a doctor who is unhappy with guidelines which say that patients should be told immediately if the test shows positive. Explain how she could use the answer to part (b) to argue that these guidelines are not appropriate.

Commentary: this question requires an understanding of conditional probability, in a context that influences the lives of many people. Screening for disease offers both costs and benefits. Benefits are obvious, and the financial costs of a screening program can be calculated. However, there are other costs that are less easy to determine. For example, when HIV first appeared, there were a number of suicides by people who turned out not to have HIV but who were told that their (first) test result was positive.

> Q2. Peter is a manager in a company which produces bottles of water. One of the machines is old and the standard deviation of the volume of water put into bottles has risen to 10.3 ml. A new machine has a standard deviation of 2 ml. Both machines have a

> setting which allows the operator to set the mean volume for the process, and the volume dispense may be assumed to be Normally distributed. The machine is being used to fill bottles which claim to contain 500 ml.
>
> When the regulator visits he tests a random sample of 10 bottles from each machine and reports a machine as defective if the mean volume of the 10 bottles is less than 500 ml.
>
> Peter wants to have only a 1% risk of a given machine being reported as defective.
>
> a) What setting would he need to use for the mean volume with the current machine to meet this condition?
>
> For the new machine the mean volume could be reduced to 501.5 ml and still have a risk of less than a 1% risk of the machine being reported as defective.
>
> b) Give one financial argument in favour of keeping the old machine and using the setting you found in part a), and one financial argument in favour of buying the new machine and setting the mean at 501.5 ml.

Commentary: this question uses realistic data to assess testing knowledge about the Normal distribution, with a context to prompt understanding of the nature of variability in production processes.

> Q3. In the past, the time, in minutes, for a particular minor medical procedure has been found to have mean 34.2 minutes and standard deviation 2.6 and can be modelled by a Normal distribution. A new method is being considered in the hope that the average time would be lower. A random sample of 50 procedures using the new method is taken and the mean time is found to be 33.5 minutes.
>
> i) Carry out a test at the 5% level of significance to see whether the mean time for the procedure has decreased.
>
> If the new method is to be adopted nursing and surgical staff will require extra training.
>
> ii) What factors should the hospital administrators take into account when deciding whether to adopt the new method?

Commentary: this question uses realistic data to assess testing knowledge about the Normal distribution, with a context to prompt consideration of whether a 'significant result' is enough to warrant a change in procedure.

Using a wider variety of data visualisations

Sutherland and Ridgway (2017) have argued that students need to be equipped with skills to interpret and critique novel data representations. A review of software to create different sorts of data visualisations (and often to analyse data) can be found at https://iase-web.org/islp/pcs/documents/Dynamic-Visualisation-Tools.pdf

Using large scale data sets

A very large number of large scale data sets are available in the public domain – some examples can be seen at https://iase-web.org/islp/pcs/documents/Data_Sources.pdf

Figure 7 presents examples from CODAP – which provides both LDS and visualisation tools.

https://codap.concord.org/releases/latest/static/dg/en/cert/index.html?url=https://concord-consortium.github.io/codap-data/SampleDocs/Social_Science/World/Human_Development_Index/Human_Development_Index.codap

Each graph was generated by drag-and-drop from the United Nations Human Development Index data set, that has been uploaded onto the CODAP website. Each point represents a country. These displays could be used to support a range of key statistical questions about definitions, data acquisition, the nature of indices, and functional relationships, as well as provoking questions about both causality and possible intervention strategies.

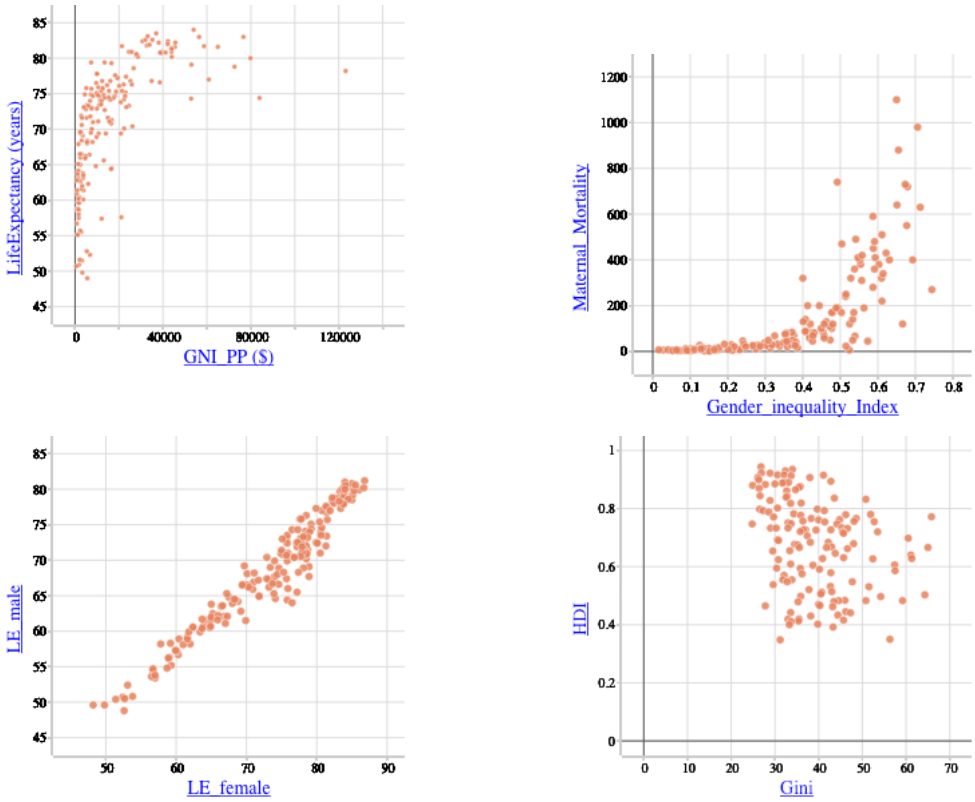


Figure 7. Data from the United Nations Human Development Index data set

7. Conclusion

We have shown that current examinations in England are not fit for the purpose of assessing or promoting competences in statistics. One reason is that the curriculum specification is ill-suited to the educational needs of students – essentially, too much is grounded in pre-computer conceptions about the nature and practice of statistics. Future revisions should take a more radical (and technology focussed) approach to the statistics curriculum.

This has been a concern for some considerable time. We argued (see Nicholson, Ridgway & McCusker, 2006; Ridgway, Nicholson & McCusker, 2007) that, even then, reasoning from data was pervasive in society and that the curriculum did not equip students with the appropriate skills to function well when working with data. We proposed that 'there was scope for substantially reducing the amount of time spent on repetitive, routine tasks such as calculations of summary statistics and graph drawing'.

There needs to be a greater emphasis on topics such as measurement, estimation, understanding evidence (e.g. in graphical displays) and reasoning with evidence – notably drawing conclusions from evidence. If LDS are to be used (which we support strongly), students should be introduced to working with real data from an early stage in the

curriculum, with the data sets increasing in scale and complexity. The curriculum content should be extended to include disaggregation of populations to allow comparisons between groups. This facilitates the exploration of the implications of policies which treated all groups uniformly, including instances of Simpson's Paradox where trends can be reversed when a population is disaggregated. In qualifications where technology and mathematical modelling are to be pervasive, it is perverse that a straight line is the only relationship between variables which is considered (imagine this constraint being applied to school science), and this should be rectified. We believe it is time to revisit the constraints on the use of technology in assessment, in particular whether it is appropriate that timed written examinations are used as the only means of assessing competence in statistics.

Fuller discussions of issues surrounding the statistics curriculum can be found in Ridgway (2016), Nicholson, Gal and Ridgway (2018), and Ridgway, Nicholson, Gal and Ridgway (2018).

References

Anscombe, F. J. (1973). Graphs in statistical analysis. *American Statistician*. *27* (1), 17-21. doi:10.1080/00031305.1973.10478966. JSTOR 2682899.

Cobb, G. W. (2015). Mere renovation is too little too late: we need to rethink our undergraduate curriculum from the ground up. *The American Statistician, 69* (4), 266–282.

Nicholson, J., Gal, I., & Ridgway, J. (2018). Understanding civic statistics: A conceptual framework and its educational applications. A product of the ProCivicStat Project. Available from: http://IASE-web.org/ISLP/PCS.

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., & McCusker, S. (2007) Teaching statistics - despite its applications. *Teaching Statistics, 29* (2), 44-48.

Porkess, R. (2012). *The future of statistics in our schools.* Available from http://www.rss.org.uk/Images/PDF/publications/rss-reports-future-statistics-schools-colleges-roger-porkess-2012.pdf.

Pullinger, J. (2014). Statistics making an impact. *Royal Statistical Society . A, 176* (4), 819-839.

Ridgway, J. (2016). Implications of the data revolution for statistics education. *International Statistical Review, 84* (3), 528-549. doi:10.1111/insr.12110.

Ridgway, J., Nicholson, J., Campos, P., & Teixeira, S. (2018). Dynamic visualisation tools: a review. A product of the ProCivicStat project. Available from http://IASE-web.org/ISLP/PCS.

Ridgway, R., Nicholson, J., Gal, I., & Ridgway, J. (2018). Understanding statistics about society: A brief framework of knowledge and skills needed to engage with civic statistics. Available from: http://iaseweb.org/icots/10/proceedings/pdfs/ICOTS10_7A1.pdf.

Sutherland, S. & Ridgway, J. (2017). Interactive visualisations and statistical literacy. *Statistics Education Research Journal, 16* (1), 26-30.

Teixeira, S., & Campos, P. (2018). Data sources. A product of the ProCivicStat Project. Available from: http://IASE-web.org/ISLP/PCS.

Appendix 1. Approaches to graphical representation

Many of the items on high-stakes examinations we have reviewed require students to demonstrate mastery of skills that have been rendered largely irrelevant by technology. Examples include drawing histograms, and drawing (often dubious) conclusions from histograms in situations where accurate answers could have been generated easily from raw data. Software can produce a range of different graphs for a given data set. Students should devote their efforts to choosing appropriate representations, and to critiquing options (for example, is the default in the software the best option for the story they want to tell?) *en route* to interpreting data and its implications. Here we provide a simple example via an exploration of what can be done with a set of real data.

Table A1 shows data giving the number of candidates gaining A level grades in Maths and English in 2015. Bar charts can be used to make comparisons, as Figure A1 shows.

Table A1. Grades awarded in A-level examinations (2015)

	A*& A	B	C	D	E	U	Total
Maths	38,660	20,304	15,019	10,198	5,934	2,596	92,711
English	16,239	24,358	27,121	13,226	2,427	335	83,706

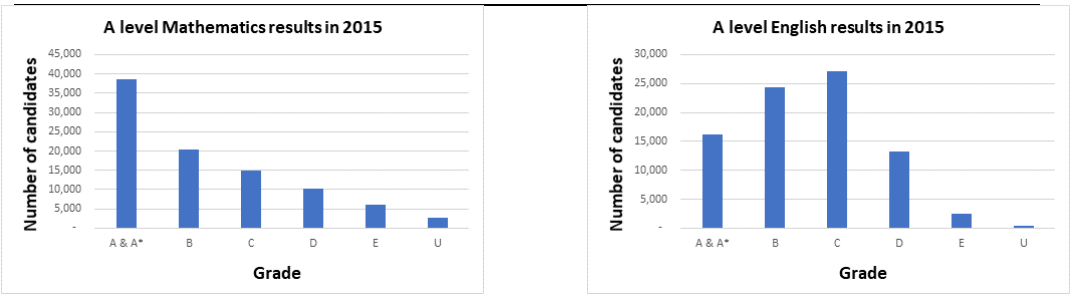


Figure A1 a, b. Bar charts for examination results in two subjects using default scales

Differences between the results for the two subjects can be seen immediately - but there is a major problem - the scale on one chart goes to 45,000 and the other only goes to 30,000. So the immediate visual impression from this autoscaling is that there appear to be many more candidates taking English than take Mathematics.

Forcing a common scale solves the problem, as Figure A2 shows. The 'cost' is extra white space at the top of one of the charts.

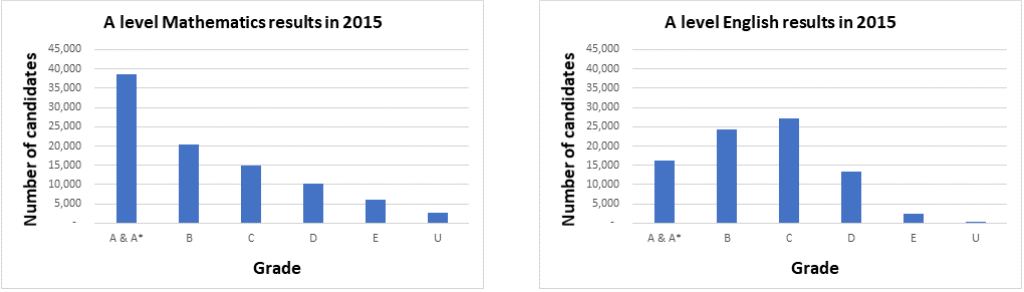


Figure A2 a, b. Bar charts for examination results in two subjects using same vertical scale

Where there is more than one set of data a comparative bar chart (or multiple bar chart) can be used to show them on the same chart, and show any similarities or differences. This immediately forces the two sets of data to be shown against the same scale. A key is needed to show what the different colour bars refer to. However, there is an option to interchange the way rows and columns are displayed – see Figure A3, a & b. In figure A3

a it is much easier to make comparisons between performance in the two subjects for each grade, while in figure A3 b it is much easier to see the dramatic difference in the profile of the distributions for the two subjects.

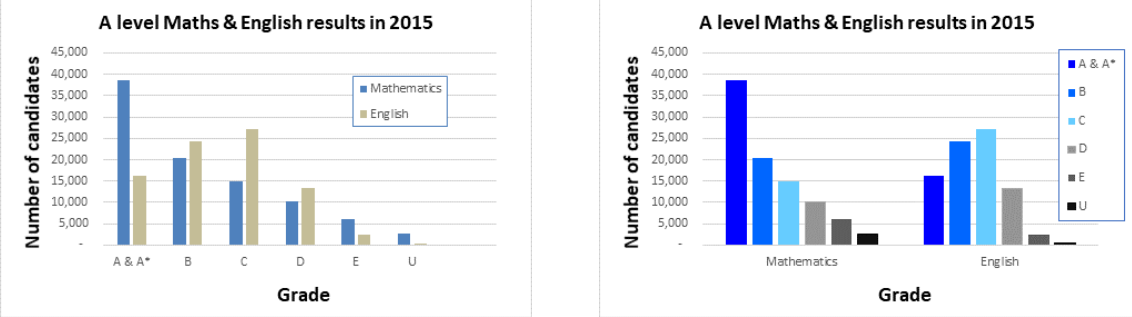


Figure A3 a, b. Comparative bar charts for the examination results in the two subjects

A stacked (or compound, or composite) bar chart can be used to show how much each grade contributes to the total. Again, a key is needed.

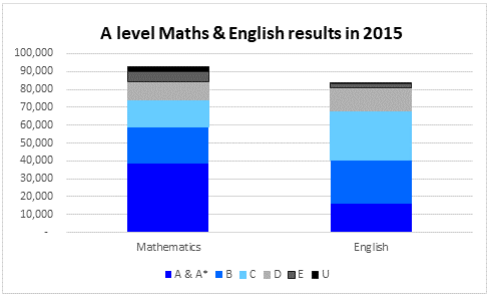
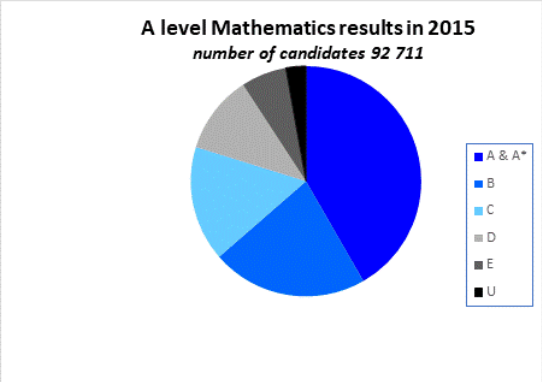


Figure A4. Stacked bar charts to compare performances in the two subjects

This graph shows clearly that more candidates took Mathematics than English, and that the proportion of A grades was much higher in Mathematics than in English.

Where the data categories are ordered, as here, the stacked bar chart lets you make cumulative comparisons as well. Here it can be seen that more candidates got at least every grade (e.g. ‘at least grade C’) in Maths than in English. (An important point to emphasise here is that stacked bar charts should never be used if the categories are not ordered).

A pie chart is the most common way to display the proportions for each group. The angle of each section of the pie chart is proportional to the size of the group it represents. While it is the most commonly used, it does not mean that it is a good representation in many of the places it is used! The pie charts shown in figures A5 a, b would be unusual because they state the size of the populations being shown, which at least gives the reader the information that there are roughly 10% more candidates taking mathematics than taking English. Standard practice is that the total is not displayed anywhere in the graphic.



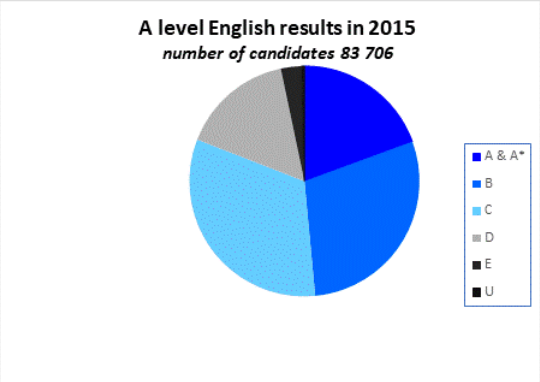


Figure A5 a, b. Pie charts to compare performances in the two subjects