Real Data AtSchool:
New Experiences with Interactive Methods of Data Exploration for Teachers

CONNOR Doreen, CROWLEY Mark & DAVIES Neville

Royal Statistical Society Centre of Statistical Education
Nottingham Trent University UK

We review our experiences with both CensusAtSchool and ExperimentsAtSchool over the last 10 years with helping teachers overcome the difficulties of coping with increasingly complex new technologies, larger and more realistic data sets while simultaneously increasing the levels of statistical literacy in students and working more conceptually. We describe our interactive web-based data investigation tool which interacts with our large datasets of real data from learners around the world and describe material and tools that we have used to help improve both teachers' pedagogy and students' learning and understanding of statistics. We believe that the next steps in statistics education should include helping teachers enjoy and understand data exploration which, in turn, will help improve students learning of statistics at school level.

1.INTRODUCTION

In 2000 we started the CensusAtSchool project hoping to " raise statistical standards and practice for teachers and pupils" (see Davies & Holmes 2000), little knowing that the project would quickly grow into the immense International enterprise it is today, offering access to real data from well over a million school children, a vast bank of curriculum enrichment material and resources for teachers and schools and offering innovative ways for learners to use their own data to compare and contrast themselves to others. It enables teachers and learners to enhance their statistical literacy and data handling skills and to improve pedagogy for teachers.

The involvement of a number of different countries increases the potential for exchange of information and widening the cultural experience. It has the added bonus of being based on websites, providing technological opportunities and is motivational for teachers and learners alike.

The ExperimentsAtSchool project developed quite naturally as an offshoot of the CensusAtSchool project based on experimental data as opposed to the survey type data of CensusAtSchool. It also helps us to utilise some of the data inputted to the online questionnaires, in particular the interactive questions such as the reaction timer, the angle estimation and the concentration game to name but a few. While teachers have to request their data back from CensusAtSchool and have it sent to them to make decisions about how it is to be used in the classroom the data from ExperimentsAtSchool is accessible directly from the website and therefore the input and retrieval of data for processing and presentation can all be done in one session. This provides opportunities for learners to interact with the data and start exploring the information it is giving them.

The increasing need to interact directly with the data led to the development of our new data tool. This tool links directly to most of the main CensusAtSchool databases and provides a way for direct dynamic interaction. It provides a way to allow intuitive graphing of random samples of data direct from the data bases.

This use of the real data presents issues and challenges that are often not present in sorted and sanitised data learners are usually given. The data also offer exciting opportunities for discussion and reflection typified by the following questions:

Which type of graph shows the data in the 'best' way?

What size sample is 'best'?

What happens if I resample the database – will I get the same picture?

How does this compare to our own data?

We will now describe some of the features of the data tool to illustrate how it can help learners to conceptualise and understand how important the processing, presentation and interpretation of real data can be.

2 DATA INVESTIGATIONS

In this section we discuss a number of ways in which the data contained in the databases stored by the RSSCSE can be used to construct meaningful investigations for teachers and learners. The RSSCSE Database Interrogation Tool has been designed to allow quick, easy and dynamic access to the large amount of data collected. We first show how the data tool can be accessed.

2.1 ACCESSING THE DATA TOOL

Figure 1 shows the opening screen of the data tool that gives access to databases constructed from the *CensusAtSchool* and *ExperimentsAtSchool* projects. We concentrate on investigating data from the first of these.

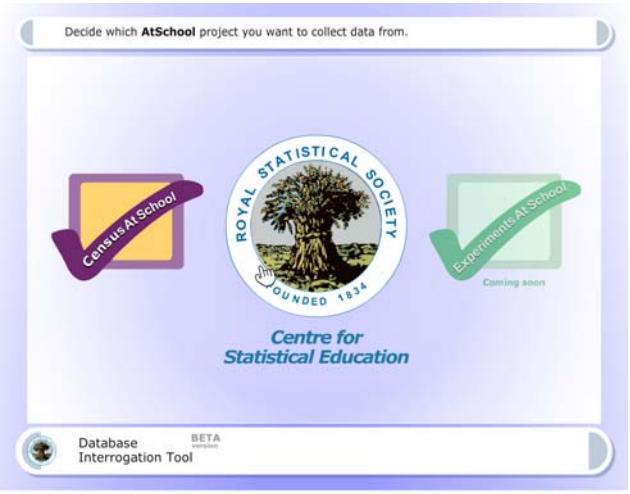

Figure 1 Opening screen for the RSSCSE Data Tool

We shall show how random samples can be obtained quickly so that graphical investigations can be carried out. Different countries have run the *CensusAtSchool* projects at different times over the last nine years and the database of responses are stored by the RSSCSE.
Users identify whether they wish to interrogate data from an individual country or from the international project data - the selection is made by clicking on the coloured flags overlaid on a map of the world. For each country selected, the number of databases available for investigation is displayed

Figure 2 Access to the Databases

2.2 ACCESSING DATA

Information about questionnaires used in each country's *CensusAtSchool* project can be accessed from a hyperlink to web pages which describe aspects of each original survey/questionnaire. Clicking on the name of a particular questionnaire or survey loads information from the corresponding database. The screen in Figure 3 shows the names of the variables in the database for the UK *CensusAtSchool* project Phase 4 in secondary schools carried out during the academic year 2003-4

The variable names are displayed as a series of colour coded tabs, showing the names of variables stored within the database as fieldnames in the tabs. Yellow tabs represent continuous variables and red tabs discrete variables. The automatic categorising facility could be the basis of a useful discussion: learners often find it difficult to decide which variables are discrete or continuous and discussing the yellow/red tab colour allocation could be beneficial.

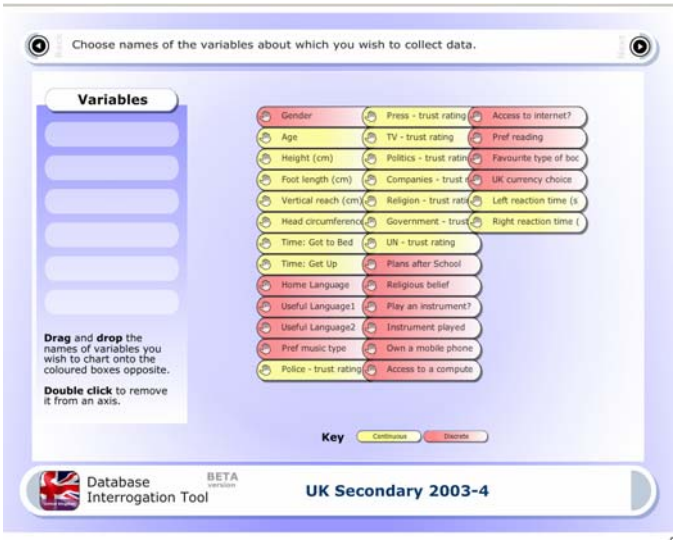

Figure 3 Details of variable type

Since teachers and learners are familiar with computer programs which use drag and drop interfaces, the tool has been based on this type of functionality. The coloured tabs can be dragged

around the screen and dropped on to the list of variables to be investigated – these are the collected variables. Users can investigate a maximum of six collected variables in one go.

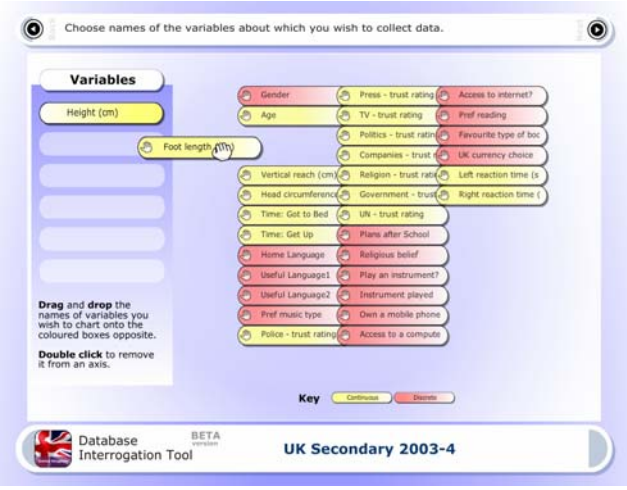

Figure 4 Drag & drop for variable selection

In the example shown in Figure 4, two variable names have been selected: Height (cm) and Foot length (cm). For particular variables a number of further options are available. The size of the sample to be collected can be varied from 50 to 250.

Sample size reflects the fact that the tool requires an internet connection; larger samples can take much longer to collect and analyse according to the user's bandwidth. In addition to sample size selection, users may collect and compare two samples and/or apply conditions to the data for selection of particular features.

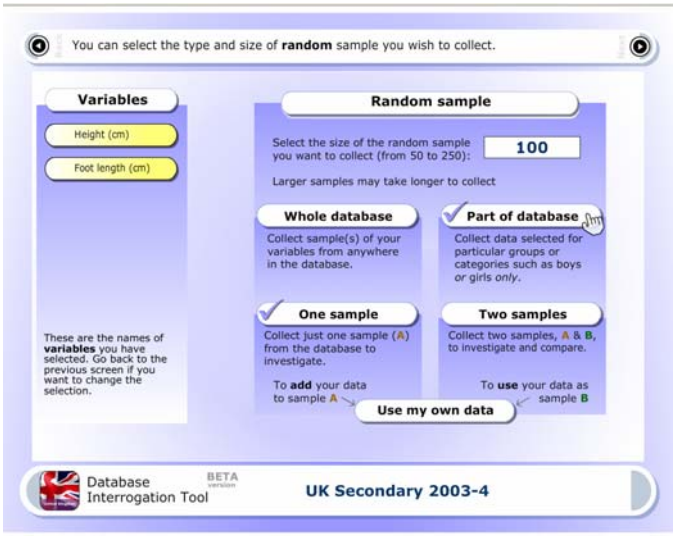

Figure 5 Selection of types of samples

In Figures 6a and 6b the ‘Part of the database’ option has been selected and a random sample is taken from a sub-set of the database, male learners aged 14.

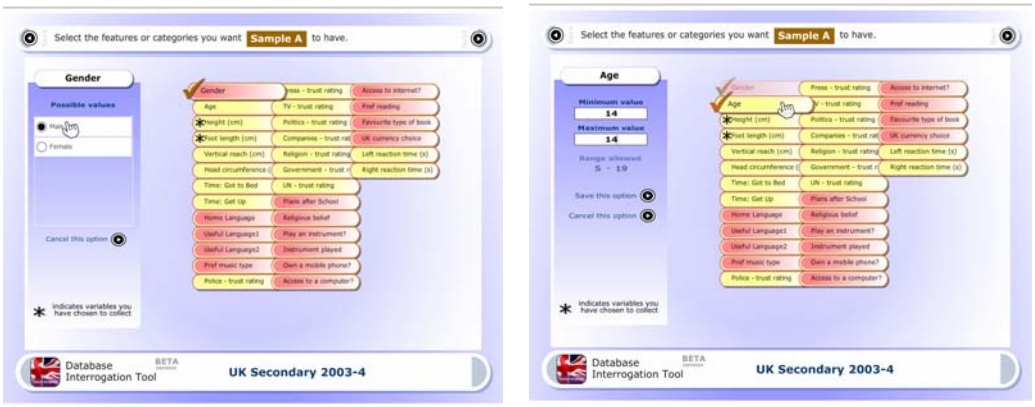

Figure 6a and Figure 6b

2.3 ACCESSING CHART OPTIONS

Once the size and, if required, other features of the sample have been selected, the data presentation window allows nine different types of graph, display and charts. These are shown in Figure 7 and comprise the following nine options.

1. Pie chart
2. Bar chart
3. Histogram
4. Box and whisker plot
5. Scatter graph
6. Tabulated data
7. Dot plot
8. 3D colour chart
9. Starplots

In this paper we illustrate the possibilities for data investigation using options 5 and 6.

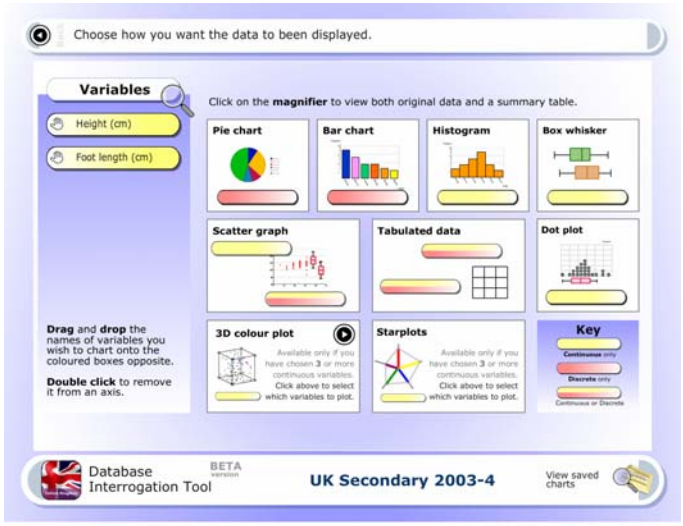

Figure 7 - Chart Selection Window

Charts are created by dragging the names of variables onto appropriately coloured tabs next to axes or table labels. The data tool prevents the user mistakenly dragging a continuous variable onto a discrete-only axis and vice versa.

Many commercial graphing and spreadsheet programmes, such as Microsoft Excel, start with the display of the raw data. They subsequently allow some basic processing and analysis, usually after highlighting particular columns of data and moving these into areas for charting. Sometimes a wizard is used to help create the displays. The data tool does not automatically tabulate the raw data, rather this is one of the options that can be chosen from the main display window shown in Figure 7. The raw data can be viewed by selecting the magnifying glass icon.

Figure 8 shows the display of the frequency chart of the heights of 14 year old male learners. Furthermore, these data can be copied to the clipboard and exported to other programs such as Microsoft Excel.

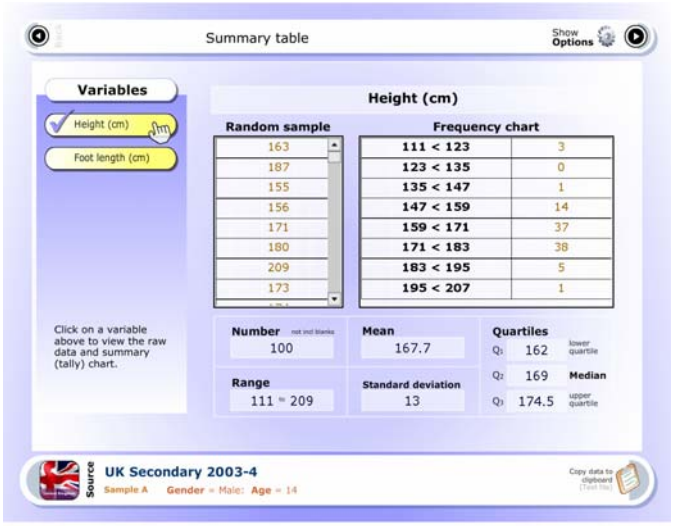

Figure 8 Raw Data Viewing

2.4 INVESTIGATING RELATIONSHIPS

Some variables have relationships between them. For example, the length of a learner's foot could be expected to increase with height. It is useful to consider the type of relationship and a scatter plot is a good way to start an investigation. In Figure 9a the continuous variables 'Height (cm)' and 'Foot length (cm)' have been dragged on to the scatter graph axes and the scatter plot is shown in Figure 9b.

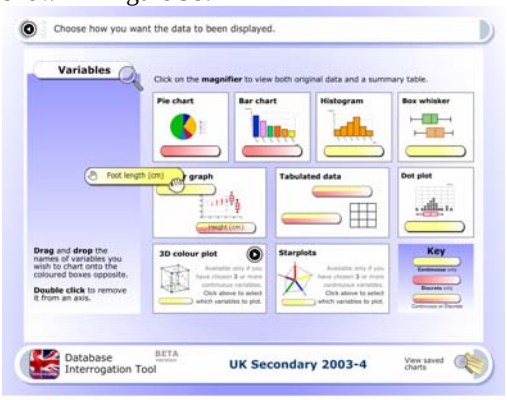

Figure 9a Dragging & dropping onto axes

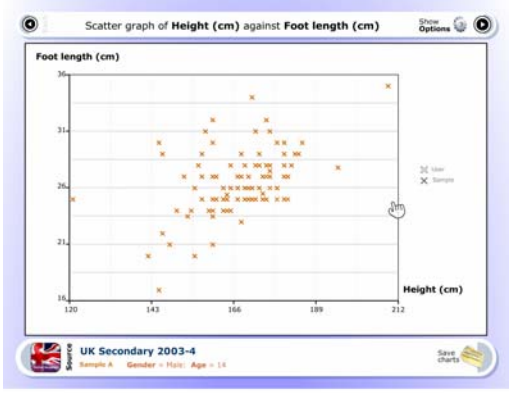

Figure 9b Scatter Graph of Foot Length and Height

Figure 9b shows the relationship between foot length and height may be approximately linear, with the amount of scatter changing as the height variable increases. There are a number of additional chart options provided by the datatool that can help a data investigation - these include adding lines of best fit and overlaying box-plots. Figure 10 shows the overlaid line of best fit for the graph of the sample of 14 year old males.

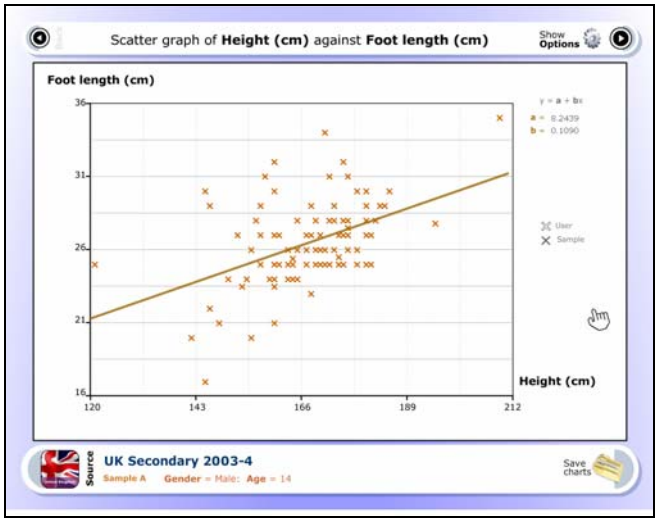

Figure 10 Scatter graph and regression line for 14 year old males

The equation of the line of best fit can be displayed, and Figure 10 shows this in the top right hand corner.

The panel below each chart displays information about the sample selected. In Figure 10 we see that the data comprise responses from males aged 14 from UK secondary schools during 2003 -04. This facility means that at every stage of an investigation using the datatool there is a constant reminder about the nature of the sample selected from the databases.

2.5 COMPARING SAMPLES

Often it is useful to compare and contrast two samples. This is possible with the data and this can be done by selecting the 'Two samples' option when choosing the type of sample(s) to use. The selection screen is displayed in Figure 11.

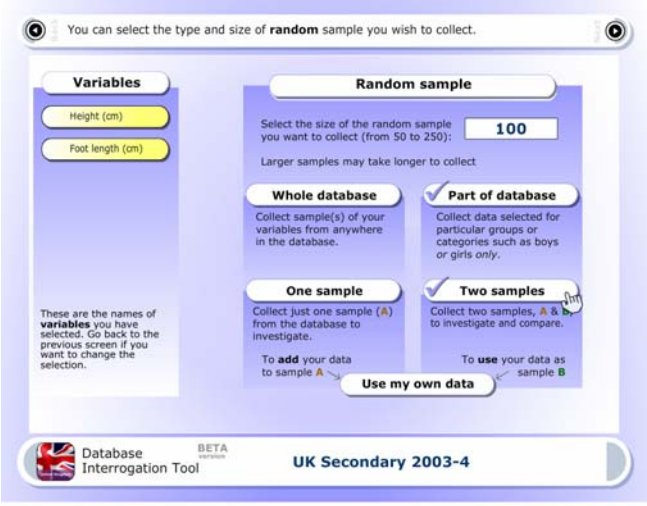

Figure 11 Two sample selection option

By selecting the 'Part of database' option, the user can select the features for the second sample, labelled 'sample B'. Figure 12 show how to collect data for males aged 14 (Sample A) and a second sample for females aged 14 (Sample B).

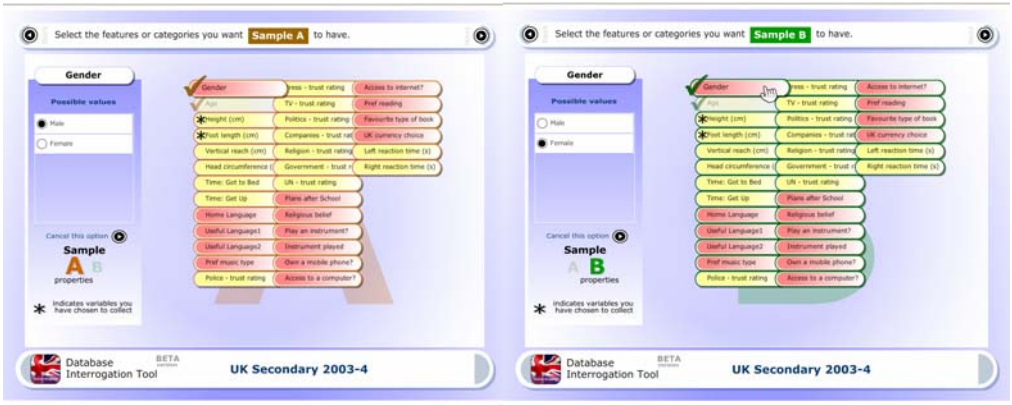

Figure 12 Selecting samples A and B

The scatter graph and fitted regression lines for the data returned for both samples is shown in Figure 13.

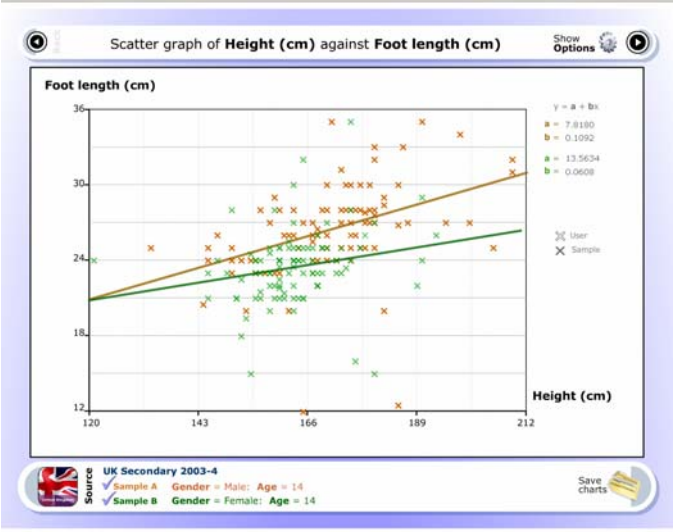

Figure 13 Scatter plot and regression lines for 14 year old males and females

In Figure 13 it can be seen that, for the two samples of males and females, the relationship between the variables *foot length* and *height* appear to be different.

It is difficult to teach the concept of natural sampling variability. For example, variation between samples can sometimes make relationships appear different, sometimes similar. One question that could be asked from looking at Figure 13 is 'What would be the appearance of the scatter graph and regression lines from other samples taken from the databases?' The data tool has a facility to answer this question through re-sampling.

2.6 RE_SAMPLING

A useful feature of the data tool is the ability to resample from the chosen database and plot the re-sampled data. The resample feature is one of the options revealed by selecting the 'Show Options' button and is shown in Figure 14.

Figure 14 Tool bar showing re-sampling option

Figures 15a, 15b and 15c show scatter graphs for three samples of 100 male and female 14 year old learners. Sample A represents males aged 14 and sample B females aged 14. Figure 15a is a plot for an original chart and two subsequent applications of the 'Resample' option are shown in Figures 15b and 15c.

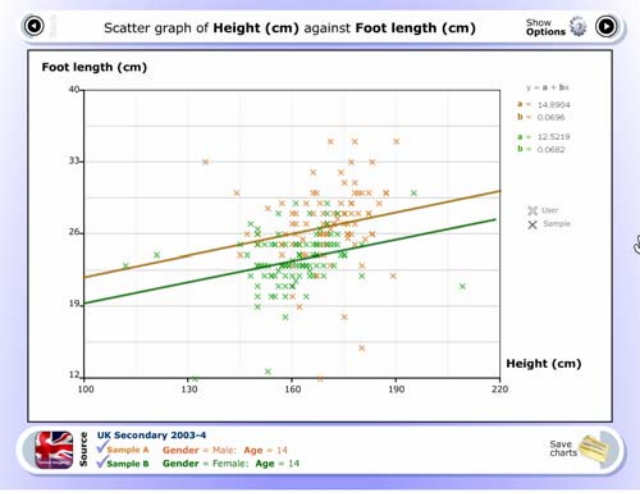

Figure 15a Second sample of 14 year old males and females

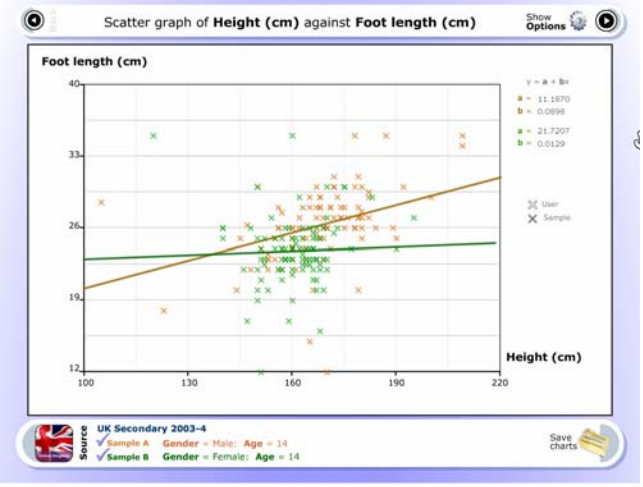

Figure 15b third sample of 14 year old males and females

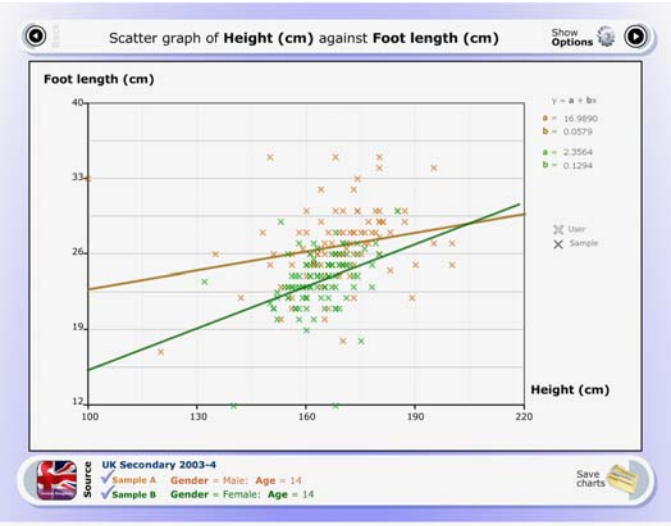

Figure 15c Fourth sample of 14 year old males and females

Figures 15a, 15b and 15c show quite clearly that successive samples of the same variables yield markedly different displays. It is fruitful to have a class discussion about why this should happen.

3. CONCLUSIONS

Formatted: Font: Italic

Formatted: Font: Italic

The capability of the data tool to show visualisations direct from the raw data in the databases is the reason it is so powerful. Learners can 'play' with the data without the need to make decisions about the most appropriate type of representation at the outset. They can use re-sampling to gain real insight into the variability within the data and easily produce very attractive and informative graphs. It is a wonderful addition to the project and the intention is to link the ExperimentsAtSchool databases as well as the remaining CensusAtSchool databases. This will further enhance the enrichment that the projects offer learners.

The increase in the use and availability of technology such as the data tool eases the burden of having to know how to produce various graphs and allows learners and teachers the enjoyment of exploring the problem itself through data interrogation. Many statistical education researchers, including Shaughnessy et al (1999), Watson et al (2000), and Garfield (2002), have been arguing for years that we need to pay much more attention to the concepts behind the techniques if we are to move forward and we believe that the data-tool is a valuable addition to moving in the right direction.

REFERENCES

Davies, N. & Holmes, P. (2000) **The Royal Statistical Society Centre for Statistical Education.** Teaching Statistics Journal, 22.1, 2-4

Garfield, J. (2002) **The Challenge of Developing Statistical Reasoning.** Journal of Statistics Education, 10 no 3.

Shaughnessy, J. M. Watson, J. M. Moritz, j. Reading, C. (1999) **There's More to Life than Centers! – Students' Conceptions of Variability**. Research Presession 77th Annual meeting of the National Council of Teachers of Mathematics. San Francisco.

Watson, J.M. & Moritz, J. B. (2000) **Developing Concepts of Sampling**. Journal for research in Mathematics education 31, 44-70.