Data, data everywhere? Opportunities and challenges in a data-rich world[1]

Nicholas Raikes Assessment Research and Development

Introduction

Everyone reading this will have heard tell of both utopian and dystopian visions for how "big data" and machine learning will change our lives. We know of the stream of data we leave whenever we use our smartphones, of the vast oceans of data held by corporate titans like Facebook and Google. We have heard how this data is the "new oil": the fuel for ever more sophisticated artificial intelligences that will change the world.

You might have come across ALTSchool (http://altschool.com). It closed some schools and refocused its strategy in 2017, but an article in *Education Week* as far back as 2016 described "lab schools" where sensors, cameras and microphones captured every physical action, every social interaction of every child every day, to supplement data gathered as the children used learning software. The vision, according to Herold, was to create:

> *Data flowing from the classroom into the cloud … a single river of information. Data scientists would then search the waters for patterns in each student's engagement level, moods, use of classroom resources, social habits, language and vocabulary use, attention span, academic performance, and more.* (2016, para. 6)

Findings from the lab schools would steer the development of the learning platform and be applied in other schools.

Despite bold visions such as this, there has yet to be a big data revolution in education – but not every claim should be dismissed as hype.

Assessment today

Big international assessment organisations like Cambridge Assessment have long held considerable amounts of data. For a typical paper-based "high-stakes" assessment, such as the International General Certificate of Secondary Education (IGCSE) or General Certificate of Education Advanced Level (GCE A Level), we know background information about most candidates, such as their date of birth, gender and school; we have their detailed marks and grades on the assessments they take with us; we know the questions they answered and who marked them; and we have their handwritten answers (as scanned digital images) and multiple-choice test responses. We use this data, for example, to give detailed information to teachers on how their students performed on the different topics tested (Figure 1), and to provide detailed information to test writers on how their questions performed, so that they can write even better questions in the future (Figure 2). More recently, we have started to use machine learning in our quality control processes. For example, we have trained a model to identify markers who are likely to be stopped due to inaccurate marking, and deployed it to monitor marks returned online and "flag" potentially poor markers for early intervention.

In this way, we can spot and fix problems sooner than we otherwise would.

1. This is an edited transcript of a presentation given at the 2018 annual conference of the International Association for Educational Assessment, in Oxford, UK. It can be viewed as a Cambridge Assessment *Research Byte* at www.youtube.com/watch?v=8_FP6YDDJ1I&t=2s

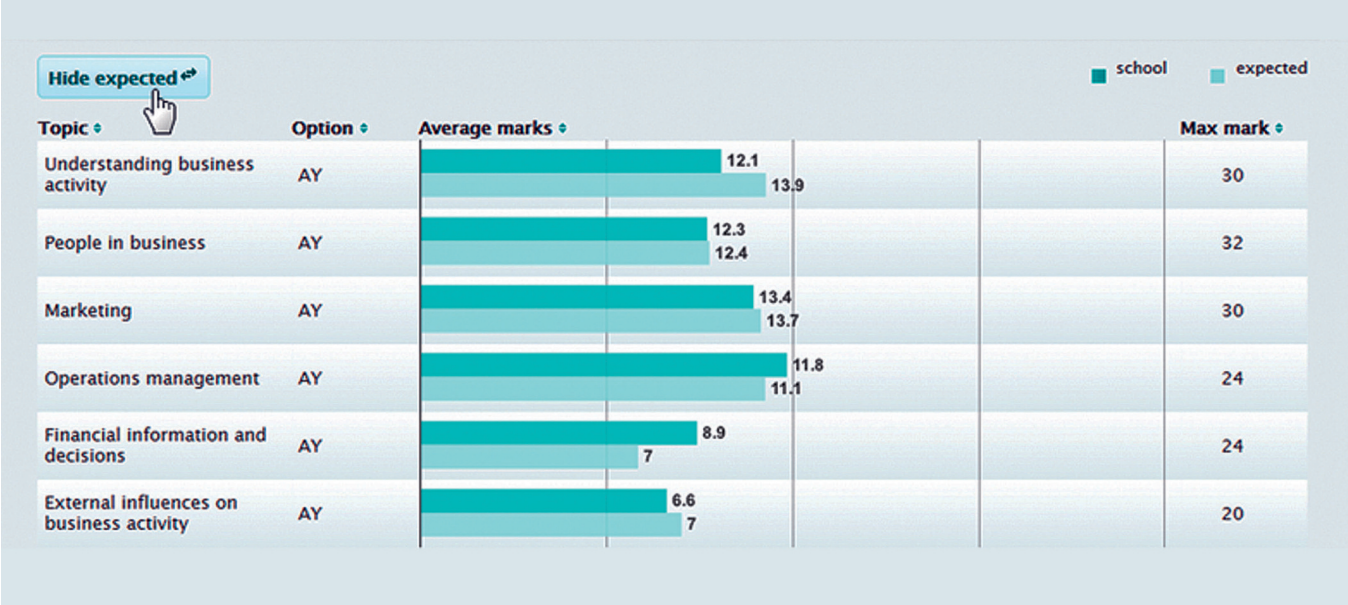

Figure 1: Results analysis for teachers
The dark bars show the average mark scored by a school's students on each topic within a qualification; the lighter bars show the average expected from a statistical model.

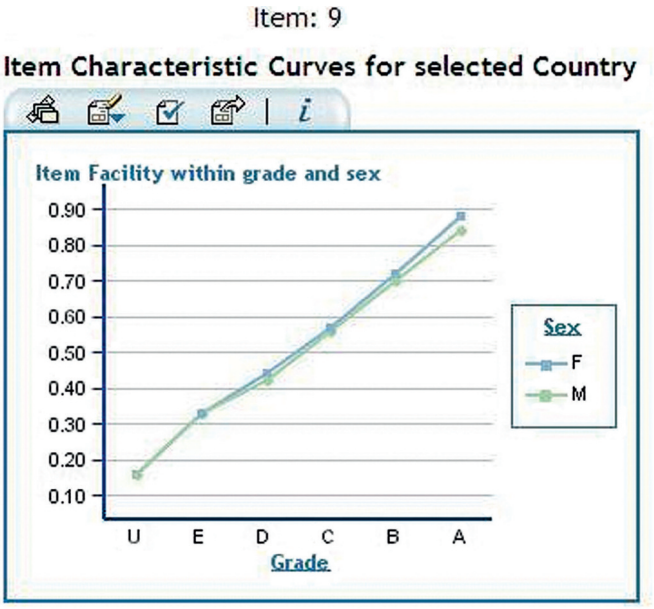

Figure 2: Question data for assessors
Detailed information is routinely provided to assessors to help with quality improvement. In this example, the chart shows how well male and female candidates with different grades overall performed on a given item.

We can even do an increasing amount with handwritten answers. Benton (2017) describes a method for spotting changes in handwriting between examinations, which could be due to an imposter sitting one (or a change of imposter). The two passages in Figure 3 were identified by the method – these were supposedly written by the same person, but this is questionable. In a personal communication to the author on September 4, 2018, Benton described how he had also demonstrated that it is possible to machine-read images of handwritten text with enough accuracy to detect some blatant cases of copying or collusion, though accuracy may depend on handwriting neatness. For example, the two passages shown in Figure 4 were detected automatically using Benton's method from amongst 18,000 others.

When text is produced digitally, we can do more with it. For example, we have operationalised computational text analytics in our question authoring and test construction processes. This allows us to screen draft exam questions automatically for any which are too similar to questions already published in text books. We also automatically screen reading passages for topic similarity in an automatically constructed, computer-delivered reading test, thereby ensuring that every student gets a variety of texts to read.

Surprisingly to many, there have been examples of automatic scoring of extended writing for around 20 years, though what works well in one context may not be applicable in all others. High-stakes tests of writing usually restrict automatic marking to providing a "second opinion" for comparison with human markers. The Cambridge Assessment English *Linguaskill* online test is used by organisations to check the English levels of individuals and groups of students, and contains a writing assessment which is automatically marked by "a series of computer algorithms that has learned how to mark test responses from a large collection of learner responses marked by expert human markers." (Cheung, Xu & Lim, 2017, p.3).

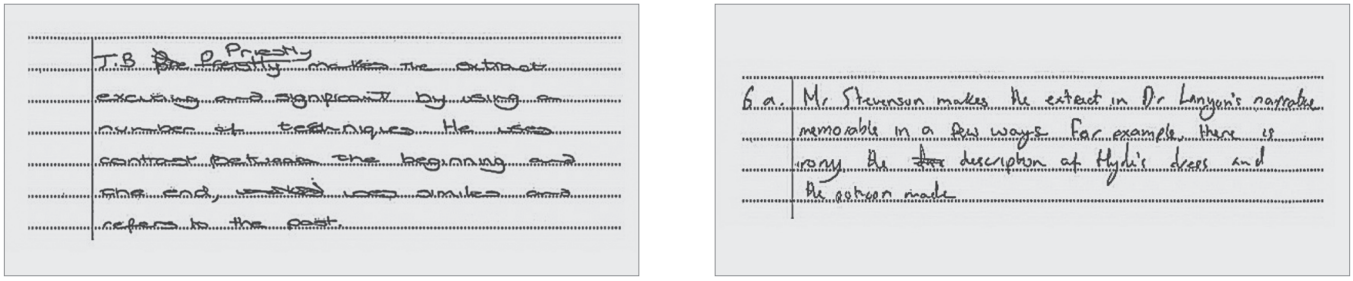

Figure 3: Malpractice detection in handwritten answers (Benton, 2017).
Both extracts were supposedly written by the same person, in different exams, but it is plausible that they are the work of different individuals. The change in handwriting was spotted automatically from changes in the median pixel density per word. Reproduced courtesy of the author.

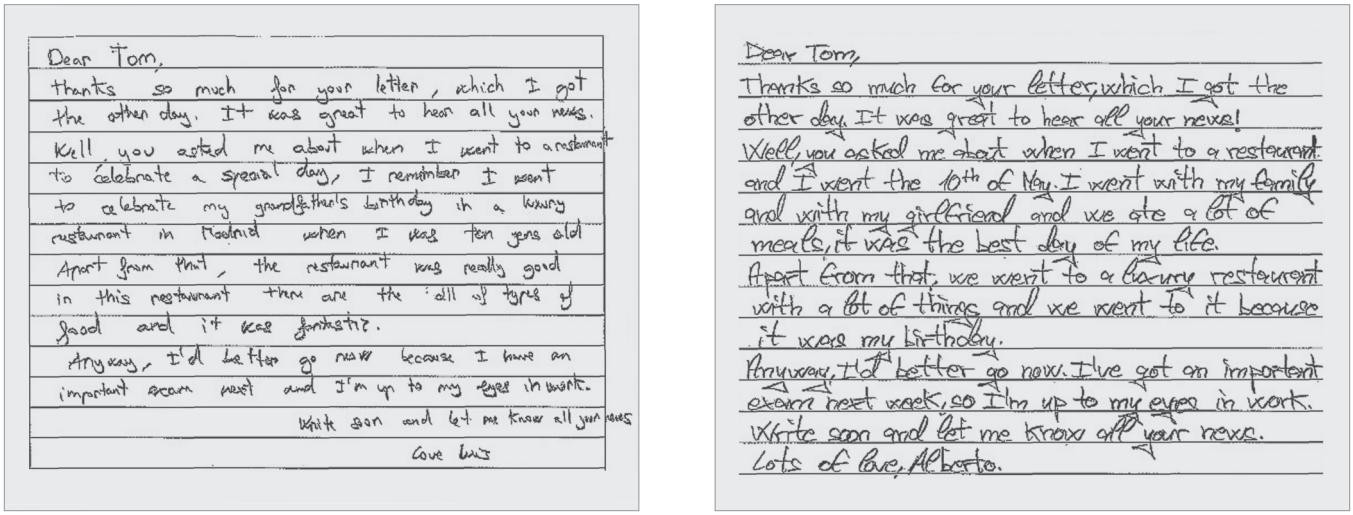

Figure 4: Malpractice detection in handwritten answers (T. C. Benton, personal communication, September 4, 2018).
Handwritten scripts were machine-read with enough accuracy to detect copying or collusion from amongst 18,000 other scripts. Reproduced courtesy of the author.

Opportunities

Let us turn now to the opportunities brought about by big data. There is no well-defined dividing line between big data and ordinary data, but big data is often considered to have three characteristics. In addition to *volume*, it has *variety* – encompassing text, video, images, log files of, for example, the key strokes and mouse clicks which students make as they engage with a computer-based test, and of the time spent focused on each task – as well as structured data, like marks and grades. This data might be streamed for analysis in almost real time at high *velocity*, which is the third characteristic of big data.

The technology and software for working with data are developing fast. Open source software such as Hadoop, Spark, R, and Python incorporate the latest advances almost as soon as they are made. Immensely powerful computing platforms can be built from relatively cheap hardware, or provided by cloud services such as Microsoft Azure and Amazon Web Services (AWS).

Machine learning has had some notable successes recently. Essentially, machine learning is statistical modelling rebranded and applied to automation. Arguably, some recent advances have had more to do with the increasing availability of data and processing power than with fundamental advances in the field of artificial intelligence. Nevertheless, these advances have potentially wide application.

In the remainder of this article, I will outline a couple of applications of big data in education, as illustrative examples of what might be possible, and finish by discussing some challenges to be overcome.

Formative assessment

The first potential application we will consider is in formative assessment; that is, assessment designed to guide learning. We would like to increase the amount of formative assessment which takes place, but high-quality questions are expensive to produce conventionally and, therefore, are scarce. However, teachers often write their own questions for use with their students. What if we provided teachers with an online platform on which they could upload their tests, and test their students, but which also made it easy for them to share and use each other's questions (items) and tests? As data accumulated, automated analytics could continuously refine estimates of item difficulty, and of how scores on one item related to scores on other items, opening the possibility that a machine-learning algorithm could be trained to categorise items, suppress bad ones, and help teachers select items and construct tests to meet their – and their students' – particular requirements. Moreover, if some of the item types were marked at first by teachers, the platform would accumulate data which could be used to train a machine-learning algorithm to mark the items automatically, thereby increasing still further the usefulness of the item bank to teachers and students.

Improved learning materials and personalised recommendations

As we collect data from more frequent testing, we will accumulate rich longitudinal data, reflecting each student's learning trajectory. We may have their work and detailed logs of what they did, as well as their marks. This would be a powerful resource for understanding learning progression and dependencies, and could be used to improve learning materials and develop better advice for teachers and students. If we pool data from learning management systems and from formative and summative assessments, we will be able to develop more intelligent adaptive learning systems. By combining the detailed information held on a student with machine-learning algorithms trained on historical data accumulated from many students, it might be possible to provide personalised recommendations automatically such as "Nick is relatively weak at algebra. This resource proved effective at raising scores for similar students".

Challenges

Most readers will by now be wondering about data protection and privacy. These are very 'hot' topics, partly because of the General Data Protection Regulation (GDPR) which came in to force across Europe in May 2018. Partly, also, because of the furore surrounding the London-based company Cambridge Analytica, and Facebook (see, for example, *The Guardian*, 2019), but also, I believe, because many people increasingly want more control over the data they produce, and of how it is used, and wish to feel confident that it is stored securely and will not be abused. Concerns are particularly acute for data about children. The case of *InBloom* is instructive (Singer, 2014). This was a project in the US with $100 million funding to create a detailed repository of educational data which would enable the kinds of application discussed above. It failed because of concerns, and then campaigns, about privacy and data protection, which snowballed until school districts and then states withdrew and the whole project collapsed. The failure highlighted the importance of establishing trust when undertaking projects such as this. This trust will depend, I believe, on having clear ethical principles and scrutiny; on being open; on communicating often and effectively to assuage concerns and inspire data subjects with the vision behind the work, encouraging them to see their data contribution as positively as volunteering or making financial donations to charity; and on gaining informed consent.

A less obvious challenge is statistical naivety. Limitations and caveats over statistical findings apply to all kinds of data, including big data. Correlation does not necessarily imply causation; data might not be representative of all people of interest, particularly if it is opportunistic; and important factors might be ignored because no data on them is easily available. Machine learning is applied statistical modelling. It is important to heed the wisdom of experienced statisticians and data scientists.

Machine learning is often described as a "black box". Models can be complex and involve many variables and levels of interaction. They are usually under the control of a data scientist, but can be hard to interpret in everyday terms. Algorithms may learn to be biased if trained on data which is biased. Microsoft's Tay chatbot was a famous early example of an algorithm skewed by training data, in that case due to pranksters feeding it extreme content on Twitter – see, for example, Lee (2016). Biases are not always so obvious, however. Sometimes algorithms are kept in black boxes for commercial reasons, their owners unwilling to be transparent about how they work.

For educational applications, we should insist on as much transparency as possible. For example, formative assessment is often described as "low stakes", but if a machine-learning algorithm gives poor advice to students, clearly their learning might be damaged, and the effects could be widespread if the algorithm is widely used. Students, like all human beings, do not always follow neat learning

progressions, making it hard to tell, for example, whether a poor performance in a short, formative test is a random aberration, or evidence of a fundamental misunderstanding and a real learning need. Developing tools of analysis and communication that can deal with this inevitable ambiguity is tricky. We should investigate the validity of machine-learning outputs, and whether they are aligned with alternative sources of evidence. And, we must evaluate the impact of data-fuelled approaches and machine learning products as they are introduced – and look for unintended consequences.

Cambridge Assessment has long been data driven. Big data, the convergence of teaching, learning and assessment, and the increasingly sophisticated operationalisation of machine learning and of data science more generally, are creating real opportunities for improving our understanding and practice of education. We should never put our faith in black boxes, however, nor introduce wide-scale change without evaluation. We must earn public trust by establishing and upholding clear ethical principles in relation to our use of data; be open; communicate continuously about what we are doing and why; inspire people with our vision and respond to their concerns; and always remember that we rely on their consent.

References

Benton, T. (2017). The clue in the dot in the 'i': Experiments in quick methods for verifying identity via handwriting. *Research Matters: A Cambridge Assessment publication, 23*, 10–16.

Cheung, K., Xu, J., & Lim, G. (2017). *Linguaskill: Writing Trial report*. Retrieved from https://www.cambridgeenglish.org/Images/466042-linguaskill-writing-trial-report.pdf

The Guardian (2019). *The Cambridge Analytica Files*. Retrieved February 8, 2019 from https://www.theguardian.com/news/series/cambridge-analytica-files

Herold, B. (2016, January 11). The Future of Big Data and Analytics in K-12 Education. *Education Week, 35*(17). Retrieved from https://www.edweek.org/ew/articles/2016/01/13/the-future-of-big-data-and-analytics.html

Lee, D. (2016, March 26). Tay: Microsoft issues apology over racist chatbot fiasco. *BBC News*. Retrieved from https://www.bbc.co.uk/news/technology-35902104

Singer, N. (2014, April 21). InBloom Student Data Repository to Close. *The New York Times*. Retrieved from https://bits.blogs.nytimes.com/2014/04/21/inbloom-student-data-repository-to-close/

Moderating artwork: Investigating judgements and cognitive processes

Lucy Chambers, Joanna Williamson Research Division and **Simon Child** Cambridge Assessment Network

(The study was completed when the third author was based in the Research Division at Cambridge Assessment)

Introduction

For the majority of standardised summative assessments in the UK, candidates will sit examinations. However, for certain practical or performance-based components, candidates will complete a non-exam assessment, which is marked by their teachers. To ensure that the standards of marking are the same across centres[1], samples of candidates' work from each centre are externally moderated. This process entails moderators, appointed and trained by awarding organisations, viewing the work and deciding whether the teachers have marked accurately and consistently. The aim of this study was to explore the cognitive processes and resources used by moderators when making judgements about artwork submitted for moderation.

The moderation method used by awarding organisations in the UK is that of inspection (see Joint Council for Qualifications[2], 2018, for a description of the moderation process). When making their judgements, moderators must consider the sample in the context of the centre as a whole, looking for trends and patterns in the marking. The moderators can make adjustments to the centre's marking, if necessary, to maintain the same marking standard across all centres. This must not be done with a view to changing the marks of individual candidates in isolation, but with a view to ensuring that the agreed standard is applied to all candidates (see Gill, 2015) for details of how centre-level mark adjustments are made).

Few studies have explicitly examined the cognitive processes involved in moderation. The only such studies that we are aware of are those of Crisp (2017) and Cuff (2017). The components under consideration in these studies involved the submission of mostly written work. The aim of this study was to investigate whether their findings hold when moderating submitted work of a very different nature, namely for Art and Design. There is little research on the marking and moderation of artwork. In fact, reviews observe that there is little detailed or technical research on assessment in art altogether (Gruber & Hobbs, 2002; Haanstra, Damen, Groenendijk, & van Boxtel, 2015; Herpin, Washington, & Li, 2011; Mason, Steers, Bedford, & McCabe, 2005).

Subject-specific research is particularly necessary for assessment in Art and Design. Assessment in Art and Design subjects is difficult: the skills involved in arts subjects are themselves complex, and furthermore "there exist many different conceptions of these skills" (Haanstra et al., 2015, p.413). Haanstra et al. go as far as to claim there is "no consensus on educational standards in the arts" (Haanstra et al., 2015, p.413). The particular demands of assessment in arts generally mean that the "forms and models of assessment particular to other areas of learning" do not transfer satisfactorily to Art and Design subjects

1. The vast majority of examination centres are schools or colleges.
2. The Joint Council for Qualifications (JCQ) is a membership organisation comprising the largest qualification providers in the UK. One of its aims is to provide common administrative arrangements for examinations.