



Cambridge
Assessment



Research *Matters*



Issue 23 / *Spring 2017*



A department of Cambridge University

Established over 150 years ago, Cambridge Assessment operates and manages the University's three exam boards and carries out leading-edge academic and operational research on assessment in education. We are a not-for-profit organisation.

Citation

This publication should be cited as:
Sutch, T., and Klir, N. (2017). Tweeting about exams: Investigating the use of social media over the summer 2016 session. *Research Matters: A Cambridge Assessment publication*, 23, 2–9.

Credits

Editorial and production management:
Karen Barden, Research Division, Cambridge Assessment
Cover image: John Foxx Images
Design: George Hammond
Print management: Canon Business Services



- 1 **Foreword** : Tim Oates, CBE
- 1 **Editorial** : Sylvia Green
- 2 **Tweeting about exams: Investigating the use of social media over the summer 2016 session** : Tom Sutch and Nicole Klir
- 10 **The clue in the dot of the 'i': Experiments in quick methods for verifying identity via handwriting** : Tom Benton
- 17 **Evaluating blended learning: Bringing the elements together** : Jessica Bowyer and Lucy Chambers
- 27 **An analysis of the effect of taking the EPQ on performance in other Level 3 qualifications** : Tim Gill
- 35 **A review of instruments for assessing complex vocational competence** : Jackie Greateorex, Martin Johnson and Victoria Coleman
- 43 **Statistics Reports** : The Research Division
- 44 **Research News** : Karen Barden
- 46 **Data Bytes** : The Data & Analytics Team
- 48 **A new look for Research Matters** : Karen Barden

If you would like to comment on any of the articles in this issue, please contact Sylvia Green – Director, Research Division. Email: researchprogrammes@cambridgeassessment.org.uk

The full issue of *Research Matters* 23 and all previous issues are available from our website: www.cambridgeassessment.org.uk/research-matters

Research *Matters* / 23

A CAMBRIDGE ASSESSMENT PUBLICATION

Foreword

A UK newspaper headline recently declared that "China uses drones to catch students cheating in exams" (*Telegraph* 2015, June 5). Education authorities in Luoyang, central China, used the latest generation of drones, which "From heights of up to 1,640 feet...will be able to home in on radio signals created by students who are using hidden earpieces to obtain the answers to exam questions...". New technology, old problem. Elsewhere in China, miniature silk booklets dating from the middle of the Qing dynasty (1644–1912) recently came to light: 160 pages containing 140,000 characters – all drawn from the fearsome Chinese civil service entrance exams – and all in a book two-and-a-half inches long and under two inches wide. Quoted in the *Telegraph* (2009, July 15), the vice chairman of the Hainan Collectors' Association stated that "The examinees had all sorts of ways of hiding these cheat sheets. They hid them inside hats, the soles of their shoes or their lunch boxes. Some sewed them into their underwear". Fast forward to this century, and some of the uses of social media around exams (the first article in this issue) stimulated the use of 'time zoned' papers. These prevent premature exposure of items, where students in one time zone have sight of and complete their exams before others around the world take the assessment. Likewise, important analysis of minutiae gives insight into candidate identity (in the second article). Cheating appears always to have been a feature of high-stakes assessment. The methods used by exam authorities need constant innovation to keep ahead of the growing catalogue of ways of cheating. And that we are doing. Trust should also be a focus of attention. Some innovation, such as the new approach to practical work in Science GCSE and GCE in England, is the result of our recognition that we have placed contradictory demands on teachers – accountability asks them to constantly improve their results, while in their role in assessing practical work they also were expected to be the objective, remote agent of the exam board. Cambridge Assessment considered this professional contradiction to be untenable for all, not least in being a threat to the important relation of trust between teachers and exam boards. This kind of structural analysis and refinement needs to accompany the development of new approaches to exam security, since mutual recognition of the importance of fair, accurate assessment is perhaps the most important objective we need to pursue. And of course, the most straightforward and robust way of doing well in an assessment is simply this: covering the syllabus deeply and comprehensively, knowing well, and doing competently the things required. Solid learning preceding accurate assessment.

Tim Oates, CBE Group Director, Assessment Research and Development

Editorial

The first three articles in this issue feature the use of technology, albeit in very different contexts. Sutch and Klir describe the collection and analysis of 6.44 million tweets from the summer 2016 UK examination series. They used real-time data from Twitter to establish the extent of exam-related tweeting, patterns over time, topics discussed, and sentiments expressed. Their research provides insights into students' perceptions and feelings about exam questions and illustrates the way candidates deal with the challenges that they face. In his article, Benton reports on research which explores the extent to which handwriting could be checked by computers to ensure that the same person has completed all components leading to a qualification. In this challenging area, the availability of scanned images resulting from online marking processes has enabled automated analysis to take place. This research experiments with some simple metrics which can identify changes in handwriting that require further scrutiny. Efficient and effective methods would be a valuable addition to the range of methods already used to discourage fraudulent practice. Bowyer and Chambers discuss the use of technology in the context of learning in their article. They provide a brief introduction to the concept of 'blended learning' and outline issues related to the implementation of blended learning programmes. Their description of existing evaluation frameworks is followed by their own framework which includes additional constructs. This is a very useful evaluative tool which can also be applied to other technology based situations.

The last two articles move away from the technology theme and focus on qualifications that assess complex skills and competence. Gill investigates whether the range of skills developed while undertaking the Extended Project Qualification might be transferrable to other qualifications taken at the same time, and might improve performance in those qualifications. The findings from this research are important in practice and lead to a number of worthwhile areas for further research. In the final article, Greateorex, Johnson and Coleman address the challenges of assessing complex vocational competence. They review the measurement qualities of checklists and Global Rating Scales in the context of assessing complex competence. In their conclusion, they provide valuable insights into the challenges of assessment in the vocational field. They succeed in providing a firm base for those working in this area to build on when deciding which tools are useful in their particular areas of assessment.

Sylvia Green Director, Research Division

Tweeting about exams: Investigating the use of social media over the summer 2016 session

Tom Sutch and Nicole Klir Research Division

Introduction

Twitter is an online social networking service, where users can post short messages (known as tweets). In general these tweets are available publicly and can be searched. As well as enabling users to 'follow' other users, thereby receiving all their tweets, Twitter also facilitates discussion among ad-hoc communities (Bruns & Burgess, 2015) by means of a hashtag: users can tag their tweet with a particular word or phrase, preceded by a hash (#) to allow users to search easily for other tweets containing that word or phrase. The list of 'trending topics' (words or phrases that are especially popular at any given time, as determined by an algorithm) presented on Twitter's front page further enhances Twitter's use for real-time discussion among disparate groups of users. Users can opt to 'retweet' any tweets that they encounter, thus passing on a copy of the tweet to their followers.

In 2013, Twitter reported 15 million monthly active users in the UK (Curtis, 2013). Younger people are more likely to use Twitter than older people: in May 2016, 33 per cent of Twitter users in Great Britain were between 15 and 24 years-old (Statista, 2016). Most analysis of social media use concentrates on adults aged 16 and over¹, but according to a survey conducted for the UK Safer Internet Centre, 37 per cent of 11 to 16-year-olds used Twitter weekly (UK Safer Internet Centre, 2015).

In recent years there has been increased tweeting by exam candidates following the end of their exams, discussing the questions on the paper. It seemed that candidates were now conducting 'exam post-mortems' in public, rather than confining them to private conversations on their way out of the exam hall (Lebus, 2016). The ability of candidates to discuss exams they have just taken with others across the country, or indeed the world, has already led awarding bodies to implement additional measures to ensure the security of their assessments. But this phenomenon has also led to coverage in national media. One notorious example from June 2015 concerned a question in an Edexcel General Certificate of Secondary Education (GCSE) Mathematics paper, and became known as 'Hannah's sweets' due to the context of the question. Candidates discussed the question and its difficulty after their exam on Twitter. Their tweets included memes², images and videos and were accompanied by the hashtag #EdexcelMaths, which began to trend on Twitter, prompting national media to run stories (for example, BBC News, 2015).

It is of interest to know the extent of tweeting about exams, and what users are saying for two reasons: first, to give some insight into the views of exam candidates into the assessments they are taking; and secondly, because stories in the national media about exam questions may indirectly shape public perception of exams and standards.

This article presents an analysis of exam-related tweets in the summer 2016 UK examination session, using real-time data from Twitter. We wanted to establish the extent of exam-related tweeting, any patterns over time, the topics being discussed, and the sentiments being expressed.

Methodological considerations

Social media data is generally available free of charge from rich *Application Programming Interfaces* (APIs) under liberal terms of use. This provides a 'free' source of data enabling extensive analysis to be carried out. Before embarking on such analysis, however, it is important to consider two issues in particular: representativeness and ethics.

Representativeness

The conversations taking place on Twitter are unlikely to come from a random subset of candidates taking exams, so there are immediate questions over how representative this data is, and what inferences can be drawn about the wider population of candidates (see Ruths & Pfeffer, 2014 for a detailed discussion of representativeness). In our context, the following are notable sources of unrepresentativeness:

- It is difficult to ascertain whether a user tweeting about GCSEs (for example) is a current candidate, relaying comments of those who are, discussing GCSEs in general, or is exploiting Twitter's currently trending topics to promote their completely unrelated tweet;
- Candidates tweeting about an exam may not be representative of all candidates. This bias can be split into two parts:
 - Twitter users may not be a random sample of candidates taking the exam (there are likely to be biases by socioeconomic background, school type and gender);
 - Conditional on being a Twitter user, candidates may be more likely to tweet about the exam if they have 'something to say': a strong opinion or emotional response, or an observation that they think would interest or amuse other Twitter users;
- Those tweeting may not be expressing their true feelings, for example, using sarcasm or humour in an attempt to connect with people, seek attention (Rui & Whinston, 2012), gain retweets, or develop a personal 'brand'.

Drawing inferences beyond the data we have towards exam candidates in general is not possible without biases being thoroughly investigated, quantified and accounted for. We do not attempt to do this in this article. However, we believe that the 'raw' data from Twitter is still useful as a source, as it shapes media coverage of exams, and thereby has a potential influence on public attitudes to exams.

1. For example, surveys on internet access carried out by the Office for National Statistics.

2. A meme is an image, video or piece of text spread from person to person. They are commonly humorous in nature and may be slightly adapted before being passed on.

Ethics

Ethics are often overlooked when analysing social media data. For example, in a systematic analysis of academic research using Twitter, Zimmer and Proferes (2014) found that only 4 per cent of studies made any mention of ethical issues or considerations in relation to the research design and data collection methods. In this case, ethics are particularly important because the subjects of interest are young people undergoing the stressful situation of taking high-stakes exams.

Henderson, Johnson, and Auld (2013) identify four ethical dilemmas associated with using social media for research purposes, focusing specifically on the educational research sphere and using children as subjects. These are:

1. **Consent in social media.** This requires a consideration of what is public and private, and there is no consensus on this. The two extremes are considering everything that is *actually* accessible as public, or only material that is *perceived* as public by participants.
2. **Traceability** resulting in a loss of confidentiality. Any verbatim text taken from Twitter is traceable (by searching for it), and researchers cannot be sure that information that is currently private is not subsequently made available at a later date.
3. **Research with children and young people.** These groups are generally considered as vulnerable subjects, necessitating greater sensitivity to consent and confidentiality. Young people have different understandings of privacy when using social media sites, and may not be considering the longer-term ramifications of posting content.
4. **Recognising and responding to illicit/reportable activities** which are evident through analysis of social media data. For example, researchers might find evidence of cyberbullying, incitement, or copyright violation.

Given these issues, we carried out our analysis at a high level by picking out numerical trends, keywords and sentiment, then aggregating, rather than looking at individual tweets. We also discarded unnecessary fields (such as real name and location) when processing the data, and did not quote any verbatim tweet content.

Data collection

We obtained data from the *Twitter Streaming API*³ which provides continuous access to real-time tweets. The particular API endpoint that we used was `statuses/filter` which returns public tweets matching one or more criteria. Access to this endpoint is free of charge, but the results returned are limited to 1 per cent of global tweet volumes.⁴ Twitter is a proprietary service, and details of the algorithms used to filter and present tweets are not published.

The only criteria we specified were on the `track` parameter, which is used to set the keywords to search for. We used the following keywords in order to capture exam-related discussion:

exam	exams	examination	examinations
resit	re-sit	GCSE	GCSEs
A Level	AS Level	A Levels	AS Levels
OCR	Edexcel	AQA	ocr exams
revise	revision	revised	revising

As well as the names of the exams, these keywords include the names and Twitter handles of the three English awarding bodies (AQA and Edexcel tweet from `@aqa` and `@edexcel` respectively; OCR (Oxford, Cambridge and RSA) tweets from `@ocr exams`).⁵ The filter is applied in such a way that searching for 'AQA' will return tweets containing handles or hashtags consisting of this word alone ('`@aqa`' or '`#aqa`') but not hashtags such as '`#aqamaths`'.

It is possible to apply a filter to return geolocated tweets posted from a specified geographical area, for example the UK. We did not opt to do this because only a minority of tweets has associated geolocation. We anticipated that the number of candidates tweeting about particular exams would be small and did not wish to restrict them further, and by inspection we found that most of our data was related to UK exams.

We used *Apache Storm* to process the stream of tweets in real time, carrying out the following additional filtering:

- Discard tweets that are in a language other than English (according to the tweet language field); retain tweets in English, and those with unspecified language;
- Require a stricter match on keywords. It is not possible to specify exact phrases to search for in the `track` parameter. For example, searching for 'A Level' will retrieve tweets that contain the word 'a' and the word 'level'; there is no requirement that they are together. We address this limitation by specifying that the words must be together, and separated by a space or a hyphen.

We saved the data at the level of individual tweets, retaining only the tweet id, text, date/time of tweet, user id, username, and (for retweets and quoted tweets) details of the original tweet.

We processed data from Saturday 14 May to Thursday 14 July 2016. This included the exam session (from Saturday 14 May to the end of June) as well as some background data afterwards to give context. In total we collected 6.44 million tweets during the exam session (from Saturday 14 May to Thursday 30 June, excluding the Half-term week beginning Monday 30 May when no exams took place).

Results

Twitter activity over time

Figure 1 is a simple plot of the number of tweets per day in our sample.⁶ Weekends are shaded to aid interpretation. A pattern within the week is immediately apparent: there are more tweets during weekdays and fewer at the weekends (especially Saturdays). There are also longer-term trends: high numbers of tweets during the start of the session in mid-May, followed by an overall decrease towards the end of the session in late June. The absolute peak was on Tuesday 17 May, when we captured 337,177 tweets. The volume of tweets was markedly lower during Half-term week (beginning Monday 30 May), during which no exams were scheduled.

3. Details are available at <https://dev.twitter.com/streaming/overview>.

4. This restriction is imposed after the filter criteria are applied.

5. We did not attempt to search for WJEC or CCEA (Council for the Curriculum, Examinations and Assessment), nor any of the associated accounts for each of the awarding bodies (for example subject-specific accounts).

6. On Friday 3 and Saturday 4 June we experienced data extraction problems, so we do not have a full set of tweets for these days.

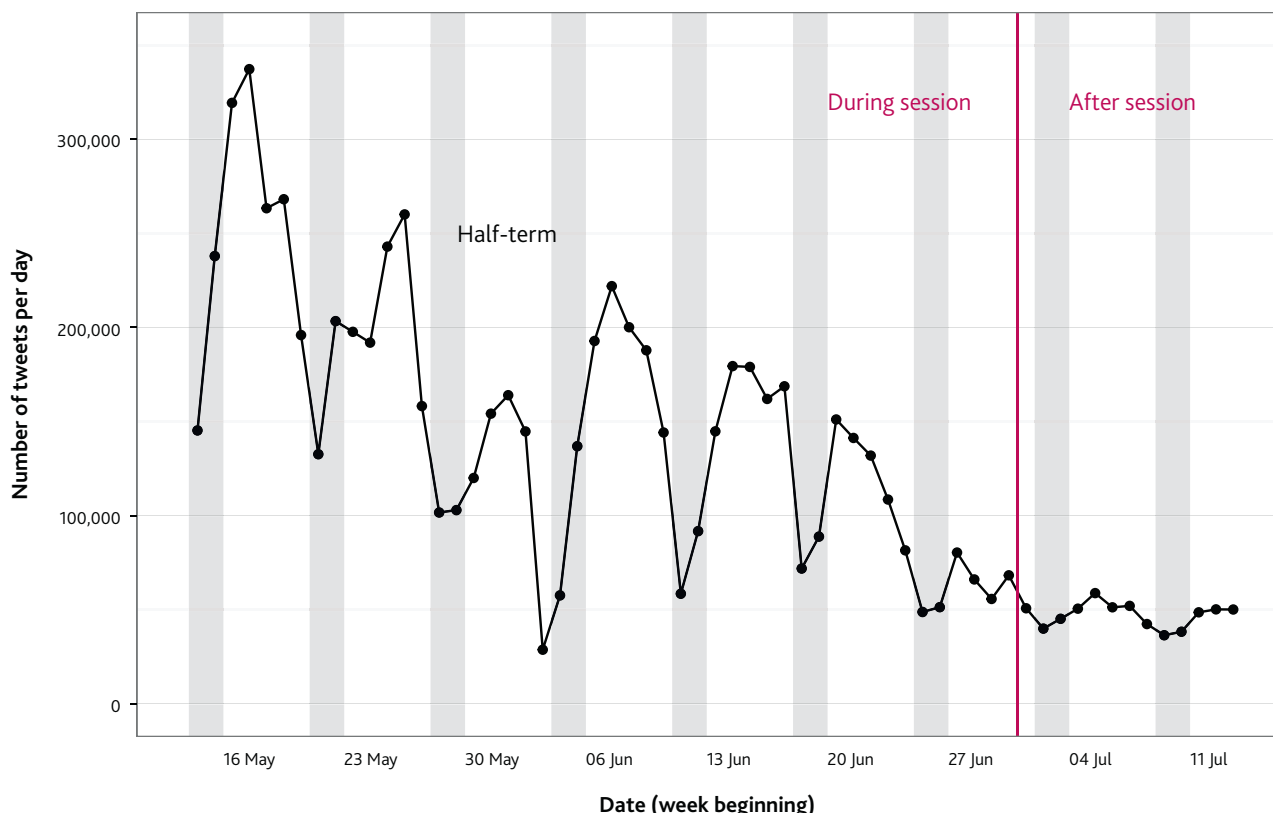


Figure 1: Number of tweets per day, 14 May to 14 July 2016 inclusive

We would expect a certain level of background 'noise'⁷ in the data, because some of our search terms may pick up tweets that are not directly related to UK school exams, and there is general discussion of exams and education on Twitter all year round. Figure 1 shows that after the end of the exam session, the number of tweets we collected was far fewer, so the extent of this background noise is fairly small.

Figure 2 shows the median number of tweets per minute, grouped by weekday.⁸ We have used the median rather than the mean because it gives us a picture of a 'typical' day; the effect of days with extremely high numbers of tweets is minimised.

To aid interpretation, we have highlighted the time periods when exams typically take place. The published starting time of examinations provided by all awarding bodies is either 9:00 a.m. or 1:30 p.m., although exam centres have flexibility to vary these start times by up to 30 minutes. The durations of exams vary; we have highlighted the most common lengths of 1 hour (dark shading) and 1 hour 30 minutes (light shading) after the start time.

The thin grey line shows the typical volume of tweets that we collected during July after the end of the exam session, for comparison. Because the volumes we collected 'in session' are markedly above this baseline, we can have confidence that we were observing genuine trends.

The chart reveals that exam-related Twitter activity follows a distinct pattern. Message volumes gradually increased in the period leading up to an exam, for example, in the evenings and the hours immediately before the scheduled start time. There was much less activity during

exam periods, but message volumes increased after the exam finished. This reinforces the 'public post-mortem' idea (Lebus, 2016).

No examinations were held at the weekends and there was a correspondingly lower number of tweets from Friday evenings onwards. However, message activity did pick up on Sundays as candidates turned their minds to the next day's exams.

Matching search terms

Figure 3 shows the number of exam-related tweets containing selected subject-related search terms on each day during the exam session, represented as a calendar for each subject. Days with more tweets are represented as darker squares. We have scaled the shading in each calendar so that the darkest shade represents the maximum number of tweets for that subject. (The entry for Latin is far lower than for Mathematics, for example, and the number of tweets is similarly lower.)

It is immediately apparent that discussion about exam subjects was concentrated on particular days. The days with the highest number of tweets generally correspond to when exams actually took place, as shown by the coloured frames, and particularly GCSE exams (which have higher entry than Advanced/Advanced Subsidiary levels [A/AS levels]).⁹ In addition, there was often a high number of tweets on the day before the exams. This can be most easily seen for the Sundays preceding Psychology AS level exams on Monday 16 May, and multiple English/English Literature exams on Monday 23 May.

For Biology, there was a large contrast between the number of tweets on Tuesday 17 May and all other days. We recorded 41,806 tweets on this day.¹⁰ This high number is due to students commenting on an AQA GCSE Biology paper which was sat on this day: these comments were picked up by many national media outlets (e.g., Espinoza, 2016).

One notable exception is the high number of tweets about History on Friday 24 June. On inspection, these tweets turned out to be related to

7. That is, tweets that do not directly relate to the exams being sat.

8. We have combined Monday to Thursday because the pattern of tweets is very similar.

9. Subject slots are collectively agreed (Joint Council for Qualifications, [JCQ] 2017, p.4), thus papers are generally timetabled together by all awarding bodies.

10. For comparison, the day with the next highest number of tweets only had 8,359.

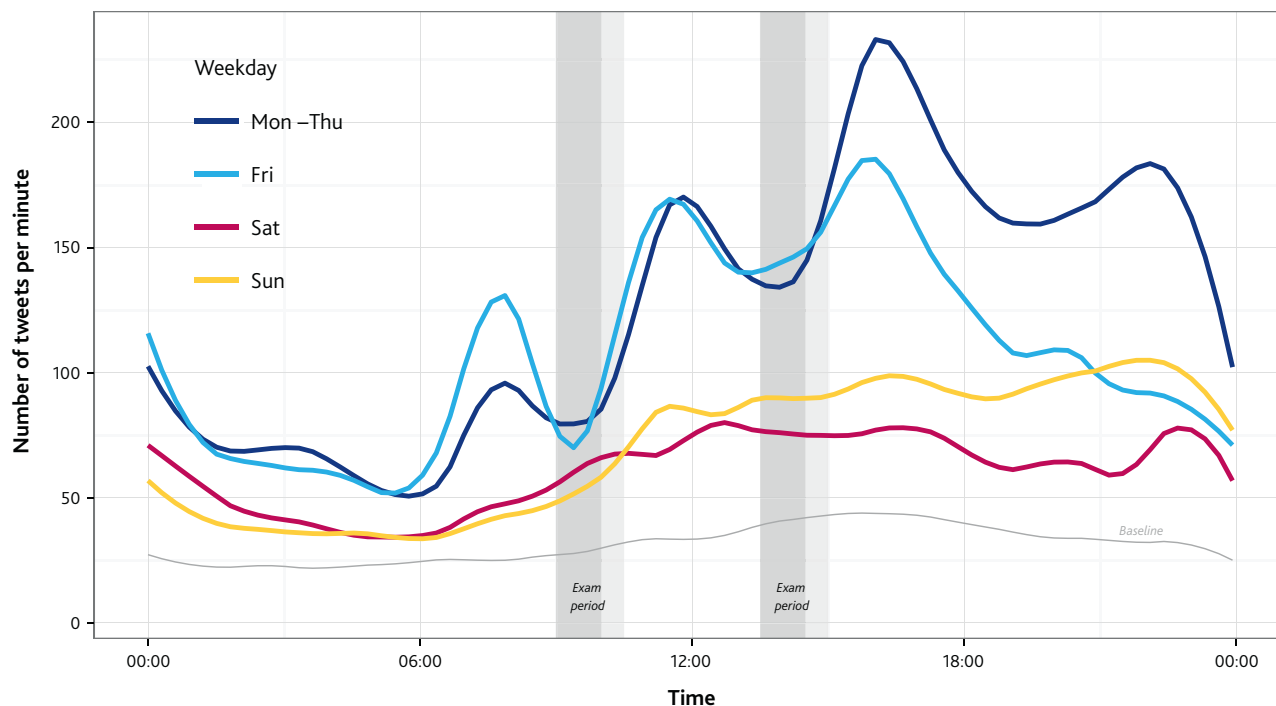


Figure 2: Tweets per time and day of week

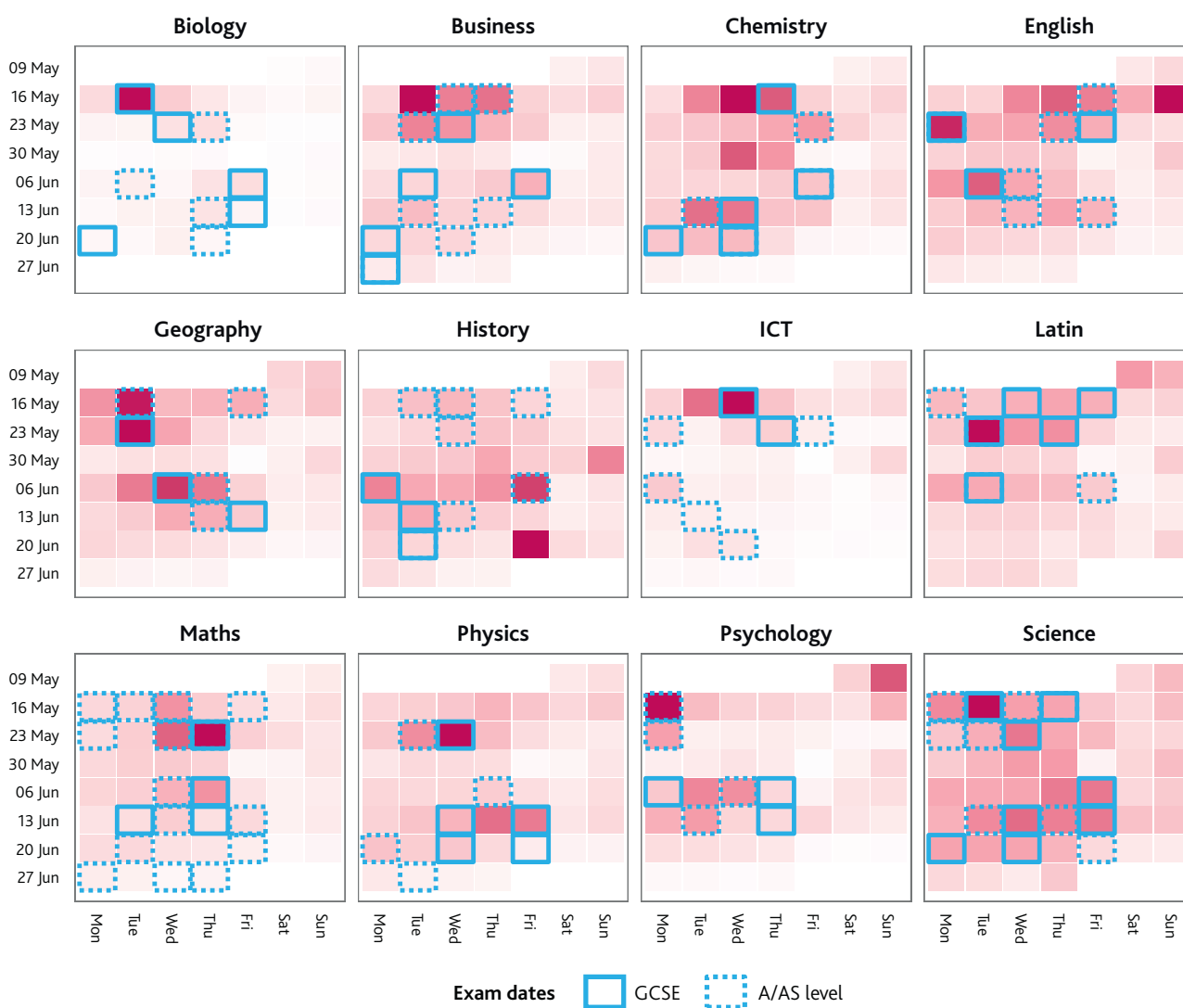


Figure 3: Calendar heatmap showing daily counts of exam-related tweets collected mentioning selected subjects, scaled by the maximum observed for each subject. (Darker squares indicate more tweets)

the result of the EU referendum held on Thursday 23 June, and subsequent political developments: many users were inventing potential GCSE History questions about these events for the future.

Tweet content

To examine the content of the tweets, we counted the occurrence of words over our corpus of tweets. Unsurprisingly the most frequently used words are those we specified as search terms. The top 20 most frequently occurring words are shown below. Those which were not specified as search terms are in bold.

exam	revising	final	school
exams	luck	GCSE	study
revision	day	week	paper
tomorrow	time	students	pass
revise	revised	examination	ur

Other than the search terms we specified, the most frequently occurring words appear to be clustered around forthcoming exams, with the fourth most common word being **tomorrow** and the seventh **luck**. By inspection of a sample of tweets containing these words, we found students were wishing each other luck, or expressing apprehension, before the exam.

To examine tweets over a period when the sample would be as broad as possible, we focused on a particular day and subject: the morning of Thursday 26 May. All the awarding bodies (including CCEA and WJEC) had timetabled a GCSE Mathematics exam for this date, and the vast majority of 16-year-olds would be sitting an exam. The scheduled start time of the exam was 9:00 a.m., and the duration of the papers ranged

from 1 hour to 2 hours. Figure 3 shows that there was a high number of tweets containing 'maths' on this day (compared to other days in the exam series).

We therefore filtered our dataset further to include tweets containing the text 'maths', sent on Wednesday 25 and Thursday 26 May, and excluding retweets. From this subset, we calculated the number of tweets featuring each word. Words were stemmed using the *SnowballC* package (Bouchet-Valat, 2014) before counting (so, for example, counts for 'revising', 'revise' and 'revised' would all be combined).

Figure 4 presents the number of tweets containing selected words before and after the exam. The time of the exam is shaded in grey.¹¹ We see that tweets referring to revision peak on the day before the exam. The words 'hope' and (especially) 'luck' dominate in the hours before the exam. Immediately after the exam, there was a surge of use of the words 'pass' and 'fail', as candidates were evaluating their performance. Words describing difficulty ('easi'¹², 'hard'), and the details of the paper ('question', 'mark', 'scheme') were also used more frequently.

Figure 5 shows the 300 most frequently occurring words used before and after the exam. Before the exam, words expressing anxiety and wishing luck are most common, whereas after the exam, words related to the exam paper and perceived performance dominate.

Sentiment via emoji

Sentiment analysis (SA) is a family of techniques for computationally determining the emotions in text. Sentiment analysis can be applied at various levels: a whole document, a sentence, or an entity such as a single phrase. In Twitter, SA techniques generally aim to determine the sentiment of the tweet as a whole. Sentiment analysis of tweets is more challenging than in other areas as tweets are limited to 140 characters (prompting users to abbreviate words and phrases), frequently include informal language, and may be on any topic in the users' interest (Giachanou & Crestani, 2016).

11. Due to the variation in duration across papers and awarding bodies, we have presented the maximum two-hour duration.

12. 'easi' is a stemmed version of 'easy'.

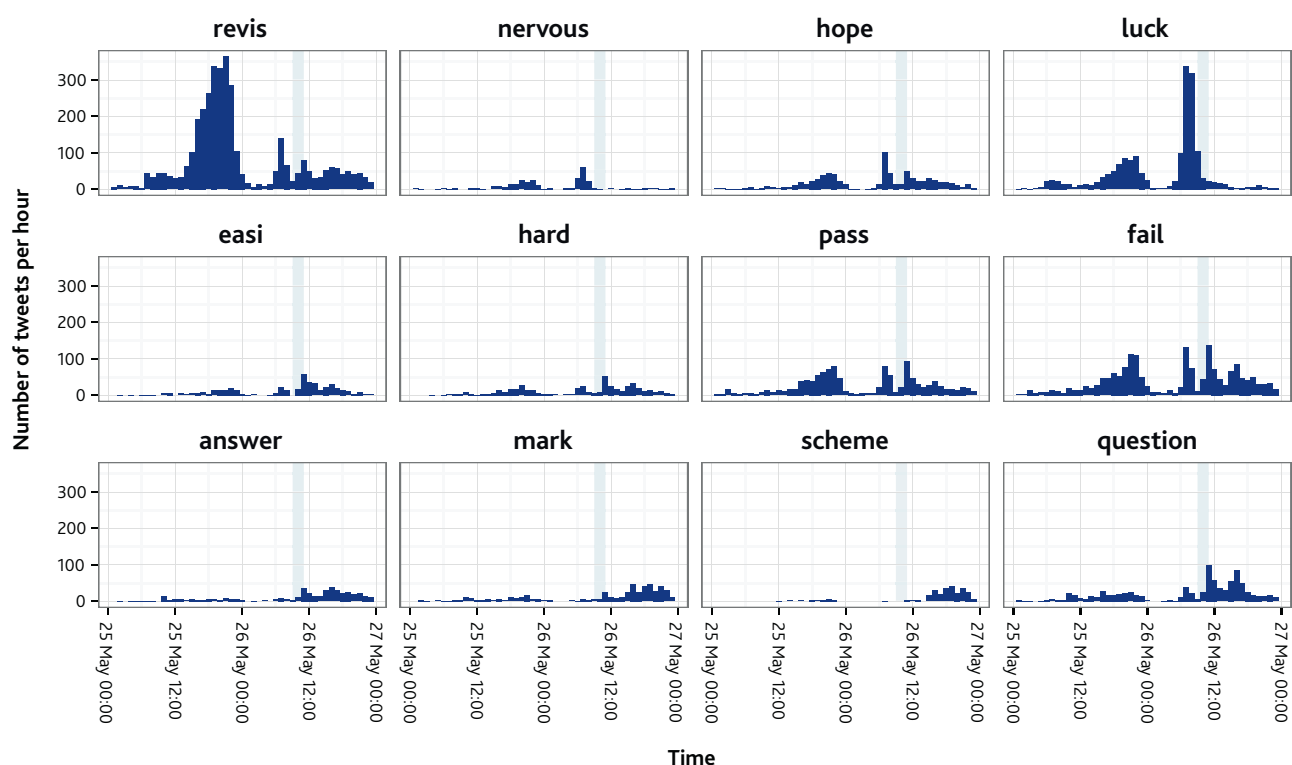


Figure 4: Use of selected words in the period around the GCSE Mathematics exam

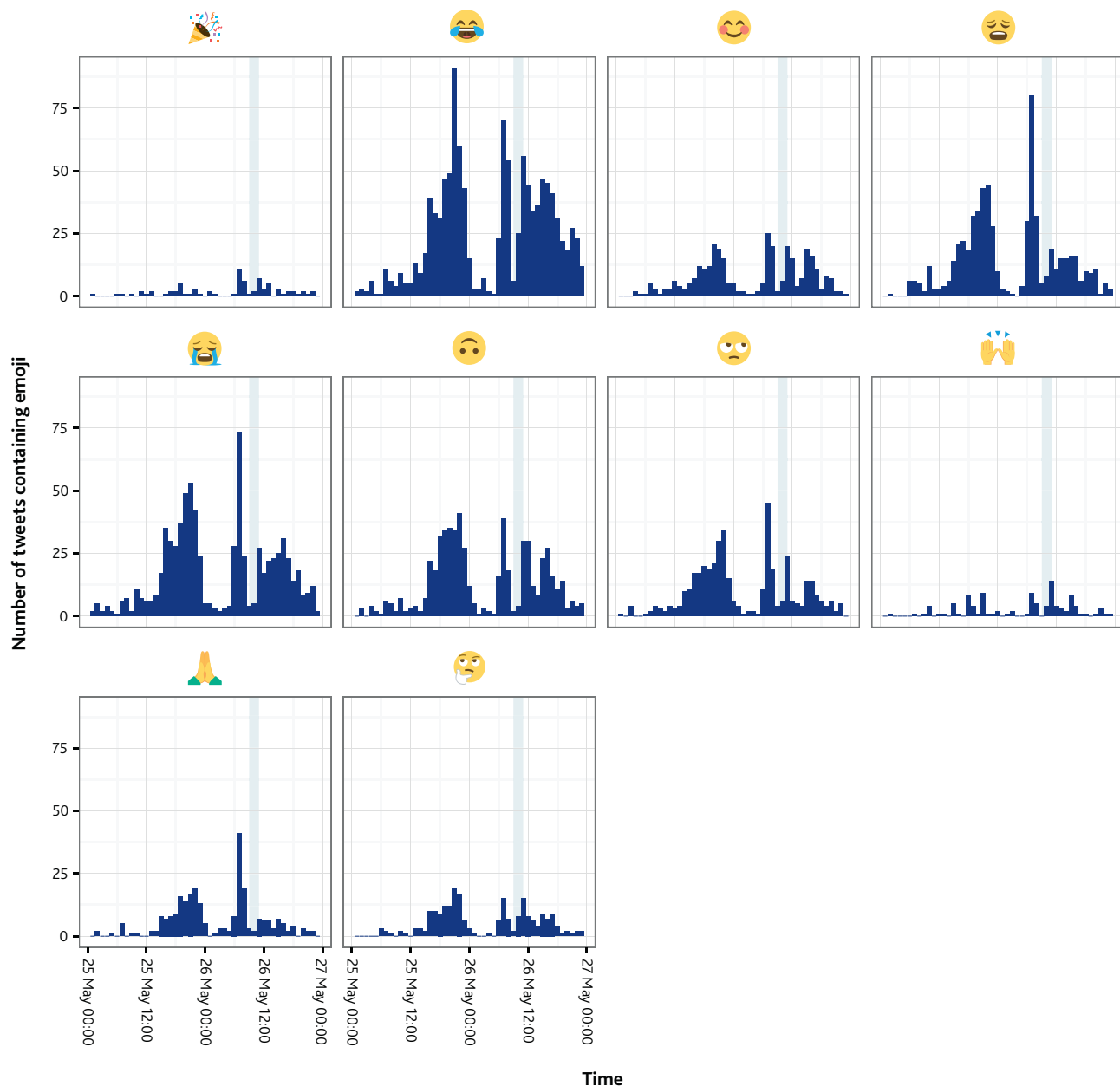


Figure 7: Maths-related tweets containing selected emoji

and 'loudly crying face' emoji over the session. By contrast, the use of the 'party popper' emoji became much more popular towards the exam session, as candidates finished their exams, especially on Fridays. For the 'person with folded hands' emoji (as if in prayer), usage was higher on weekdays when exams were taking place, but there was no evidence of any other trends.

However, mapping overall sentiment in this way does not reveal the full picture, due to the number and frequency of exams throughout the session: any positive sentiment expressed by students after finishing one exam would be counteracted by students expressing nerves and apprehension before a forthcoming exam in the following days.

In addition, we decided to look at sentiment (as measured by emoji) over the GCSE Mathematics example shown earlier, and for the same subset of tweets containing 'maths' on Wednesday 25 and Thursday

26 May. The numbers of tweets containing each of the top 10 emoji (in Table 2) for this period are shown in Figure 7. The most notable pattern here is that several emoji had surges of popularity just before the exam, for example 'person with folded hands' (as if in prayer) and 'weary face'. Immediately after the exam there were spikes in popularity for several emoji, including 'face with tears of joy' and 'loudly crying face'. Overall, there was more use of emoji in the evening before the exam, rather than afterwards.

Discussion

In this article we have described the collection and analysis of 6.44 million tweets from the summer 2016 examination session in the UK.

We found that there were more exam-related tweets at the beginning of the session than at the end. This may be because the JCQ seeks to timetable large-entry subjects as early as possible to facilitate marking (JCQ, 2017, p.5). It may also be that students are more excited at the beginning of the session for their first few exams, whereas towards the end they are becoming more accustomed to the rhythms of the exam session. We looked at the sentiments expressed in these tweets using a simple emoji approach, and found a general decrease in the use of the most common emoji ('face with tears of joy' and 'loudly crying face') over the session, perhaps reflecting such a decline in excitement.

By looking for subject-related search terms within the tweets we gathered, we identified that students were more likely to tweet about exams on the day of their exam, or in some cases the day before. We found that during weekdays there was a surge in exam-related tweets immediately before and after exam times, with peak time being the late afternoon when the day's exams were over.

From our investigations of the tweets relating to the GCSE Mathematics exams on Thursday 26 May, we found that it was possible to identify several different phases based on the words used in tweets: on the day before the exam, the dominant topic was revision. Immediately before the exam, the words 'luck' and 'hope' featured prominently; whereas after the exam there was talk of difficulty, and features of the paper (questions and mark schemes) along with an emotional response. When we looked at the sentiments expressed via emoji, we found a similar pattern, although perhaps more negative feeling before the exam than was expressed with words.

Limitations

Twitter data, and social media data in general, is an exciting data source, readily available, which offers the potential to get feedback from a large number of students taking assessments. However, this data should be used with extreme caution, as it is unlikely to be representative of students' views.

The availability and ease of extracting this data may lead us to forget that this is ultimately personal data, collected from young people undergoing stressful high-stakes exams, and ethical issues should be considered before using it.

While social media data offers the opportunity to investigate and understand candidates' views on their experience of exams, and even certain questions, actually using the information gained is fraught with risk. As discussed earlier in this article, discussions on Twitter are unlikely to be representative of students' views. Additionally, paying too much heed to comments on social media could limit awarding bodies' capacity to develop innovative items and assessments that test the full range of candidates' skills (see comments by Professor Rob Coe, cited in Busby, 2016). However, the fact that social media activity does now influence articles in the mainstream national media means that Twitter comment may affect the general public's view of exams and their standards.

References

- BBC News. (2015, June 5). *Tricky GCSE maths exam sees pupils take to Twitter*. Retrieved from <http://www.bbc.co.uk/news/education-33017299>
- Bouchet-Valat, M. (2014). *SnowballC: Snowball stemmers based on the C libstemmer UTF-8 library*. Retrieved from <http://CRAN.R-project.org/package=SnowballC>
- Bruns, A., & Burgess, J. (2015). Twitter Hashtags from Ad Hoc to Calculated Publics. In N. Rambukkana (Ed.), *Hashtag publics: The power and politics of discursive networks* (pp.13–28). New York: Peter Lang. Retrieved from <http://snurb.info/files/2015/Twitter%20Hashtags%20from%20Ad%20Hoc%20to%20Calculated%20Publics.pdf>
- Busby, E. (2016, October 21). Social media fears lead to 'predictable' exam papers. *TES*, 12–13.
- Curtis, S. (2013, September 6). Twitter claims 15m active users in the UK. *The Telegraph*. Retrieved from <http://www.telegraph.co.uk/technology/twitter/10291360/Twitter-claims-15m-active-users-in-the-UK.html>
- Emoji Keyboard. (2016). In *EmojiOne™ Version 3.0*. Retrieved January 19, 2017, from <http://emojione.com/>
- Espinoza, J. (2016, May 17). Students left 'fuming' over GCSE biology exam that contained questions about drunk 15-year-olds. *The Telegraph*. Retrieved from <http://www.telegraph.co.uk/education/2016/05/17/students-left-fuming-over-gcse-biology-exam-that-contained-quest/>
- Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of Twitter sentiment analysis methods. *ACM Computing Surveys*, 49(2), Article 28, 28–41. <http://doi.org/10.1145/2938640>
- Henderson, M., Johnson, N. F., & Auld, G. (2013). Silences of ethical practice: Dilemmas for researchers using social media. *Educational Research and Evaluation*, 19(6), 546–560. <http://doi.org/10.1080/13803611.2013.805656>
- Hogenboom, A., Bal, D., Frasincar, F., Bal, M., de Jong, F., & Kaymak, U. (2013). Exploiting emoticons in sentiment analysis. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing* (pp.703–710). New York, NY, USA: ACM. <http://doi.org/10.1145/2480362.2480498>
- Joint Council for Qualifications. (2017). *Construction of the common timetable: FSMQ, GCE and GCSE qualifications*. Retrieved from <http://www.jcq.org.uk/exams-office/key-dates-and-timetables/construction-of-the-common-timetable%E2%80%9393fsmq-gce-and-gcse-qualifications>
- Lebus, S. (2016). *Summer Series 2016 – Postmortems, petitions and practicalities* [Blog post]. Retrieved from <http://cambridgeassessment.org.uk/blog/summer-series-2016-postmortems-petitions-and-practicalities/>
- Rui, H., & Whinston, A. (2012). Information or attention? An empirical study of user contribution on Twitter. *Information Systems and e-Business Management*, 10(3), 309–324. <http://doi.org/10.1007/s10257-011-0164-6>
- Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, 346(6213), 1063–1064. <http://doi.org/10.1126/science.346.6213.1063>
- Statista. (2016). *Age distribution of Twitter users in Great Britain from May 2013 to May 2016*. Retrieved from <https://www.statista.com/statistics/278320/age-distribution-of-twitter-users-in-great-britain/>
- Twitter Streaming API. (n.d.). In *Twitter Development Documentation*. Retrieved January 19, 2017, from <https://dev.twitter.com/streaming/overview>
- UK Safer Internet Centre. (2015). *Friendship in a Digital Age*. Retrieved from <http://www.saferinternet.org.uk/safer-internet-day/sid-2015/friendship-digital-age-new-report-launched>
- Zimmer, M., & Proferes, N. J. (2014). A topology of Twitter research: Disciplines, methods, and ethics. *Aslib Journal of Information Management*, 66(3), 250–261. <http://doi.org/10.1108/AJIM-09-2013-0083>

The clue in the dot of the 'i': Experiments in quick methods for verifying identity via handwriting

Tom Benton Research Division

Summary

This article demonstrates some simple and quick techniques for comparing the style of handwriting between two examinations. This could potentially be a useful way of checking that the same person has taken all of the different components leading to a qualification, and form one part of the effort to ensure qualifications are only awarded to those candidates that have personally completed the necessary assessments. The advantage of this form of identity checking is that it is based upon data (in the form of images) that is already routinely stored as part of the process of on-screen marking. This article shows that some simple metrics can quickly identify candidates whose handwriting shows a suspicious degree of change between occasions. However, close scrutiny of some of these scripts provides some reasons for caution in assuming that all cases of changing handwriting represent the presence of imposters. Some cases of apparently different handwriting also include aspects that indicate that they may come from the same author. In other cases, the style of handwriting may change even within the same examination response.

Introduction

In order for assessments to be any use at all, it is crucial that they are taken by the same people to whom results will be issued. As such, we need measures to discourage any attempts at malpractice by one person completing an assessment on behalf of another. Reports of such forms of cheating are currently extremely rare in the UK; however, they are frequently reported in other countries and it is important that we should be prepared for the possibility of this type of cheating.

There are many possible ways of checking the identity of candidates. For example, for certain assessments internationally, candidates are required to take identification documents with them to the exam centre in order to be permitted to take the exam. As an alternative, for some tests produced by the exam board Cambridge English Language Assessment, test day photos are taken so that users of examination results are able to verify for themselves the identity of the person who actually took the assessment. However, in addition to such checks, there may be value in examining the handwriting used within assessments to verify that all of the different elements of a qualification are being taken by the same individual.

The advantage of using handwriting for identity verification is that it is a source of information that is already freely available to exam boards. The vast majority of Cambridge Assessment's examinations are taken using pen and paper and, furthermore, due to the rise of on-screen marking, scanned images of most candidates' scripts are already stored within our systems. Thus, it is theoretically possible for us to examine the handwriting used across all of the assessments taken by an individual to

help reassure ourselves that the correct individual is receiving credit for his or her work.

Manually checking handwriting from different assessments against one another would be both laborious and expensive in terms of labour costs. For this reason, the aim of this research project was to begin to explore the extent to which such a process could be automated by computers.

Automatic handwriting recognition is a widely researched area (Dolega, Agam, & Argamon, 2008) with a wide variety of available algorithms. However, many of these algorithms are slow – requiring detailed tracing of the strokes used to form each of the words and letters on a page and could not be quickly applied to the thousands of digital images we hold. Instead, this project looks for whether there are any metrics of handwriting that are relatively quick to calculate which could provide a reasonable indicator of whether the author of two separate pieces of handwritten text was the same.

Source of images

The images for analysis were extracted from Cambridge Assessment's Digital Script Repository (DSR). Since this is the first time we have undertaken analysis of this kind, a relatively simple example was chosen. Two compulsory Higher Tier papers, taken two days apart as part of a General Certificate of Secondary Education (GCSE) in English Literature in June 2014, were chosen for use in the analysis. Throughout this article, the two papers will be referred to as 'Unit A' and 'Unit B'. Both examinations required candidates to provide essay-type responses written on lined paper. Excluding the front and back covers of the script, for the vast majority of candidates 6 scanned pages were available from Unit A and 14 from Unit B, although it was rare for candidates to actually write on all of the available pages.

The consistent format of responses between the two assessments simplified the process of analysis. The aim of the project was to develop some simple measures for the style of handwriting and explore the extent to which such metrics remain stable between different examination occasions. Metrics that are highly stable between occasions might be useful for verifying that the same person has taken each examination.

Software, methodology and metrics

All of the analysis within this article was undertaken using the free statistical software package R version 3.2.2 (R Core Team, 2015). The majority of the work of reading, segmenting and manipulating images was done using the package *EBImage* (Pau, Fuchs, Skylar, Boutros, & Huber, 2010).

Pre-processing of images

Before analysis of handwriting can begin, a few pre-processing steps are necessary. The key steps of this process are shown in Figure 1. For reasons of space, the images in Figure 1 are restricted to a portion of text at the top of one page of a candidate's response.

The top left-hand of this image gives an example of what (part of) a single page of a candidate's response might look like before any pre-processing has been applied. To begin with, the full-page image is read into R as a grayscale matrix. The data matrix has one value for each of the 2300×1620 pixels in the image and, as it is standard for data representing images, higher values are given to whiter sections of the image and lower values to the blackest sections – that is the sections where there is actual writing.

The first step of pre-processing is to attempt to distinguish shapes that represent actual handwriting from those that represent margins, the typed text of the question, or dotted lines. This task requires a number of steps.

To begin with, 5 per cent of the original image is removed on both the right and the left. This is done to remove dark black lines that may be created at the edge of the image as part of the scanning process. Next, the image is converted from grayscale to black and white using *Otsu's method* (https://en.wikipedia.org/wiki/Otsu's_method). Now each pixel in the image is either represented by a 0 (white) or a 1 (black). Next, we break the image into sections of joined up black pixels. Due to the medical context of the software (which treats white sections as indicating the presence of something and black sections as absence), it is necessary to take the negative of the image before doing this. The identified separate sections of joined up pixels are shown in different colours in the top right-hand corner of Figure 1.

The size of each segmented section in the top right-hand image can be used to identify two sets of items of interest: dots from the dotted lines

and written words and letters. Some experimentation revealed that the dots printed within dotted lines on the exam paper tended to contain between 10 and 55 pixels. Using this rule of thumb, we could count the number of pixels within such dots in each row of the image. Rows where between 80 and 200 pixels were within these identified 'dots' were deemed likely to represent a dotted line¹. Thus, we could identify the first and last such rows in the matrix as likely representing the top and bottom dotted lines, and restrict the matrix to writing between these two only. One downside of this approach was that if the candidate wrote on top of the first dotted line then this text was lost. In addition any text below the final dotted line is also lost.

Some experimentation showed that most sections that represented handwriting (that is, words or elements of words) tended to contain between 60 and 3,000 pixels. Any joined groups of pixels outside of this range were set to be white as they were unlikely to represent writing. However, this often meant that the dots of handwritten 'i's or 'j's were deleted.

Applying the steps above led to images of the type shown in the bottom left-hand corner of Figure 1. Some simple metrics of handwriting, to be described later, were calculated based purely upon this image.

However, one problem with using the image so far was that the metrics of handwriting may be affected by the thickness of the pen used by the candidate. This was addressed by using a crude form of 'thinning'. Thinning is the process of trying to find a skeleton form of any given shape that is only one pixel wide at any point but that preserves the essence of the shape. Many algorithms have been proposed for this procedure (see Lam, Lee, & Suen, 1992). However, the formal approaches

1. This obviously ignores the height of the 'dots'. We are just looking at the number of pixels within the dots, within each row of the matrix.

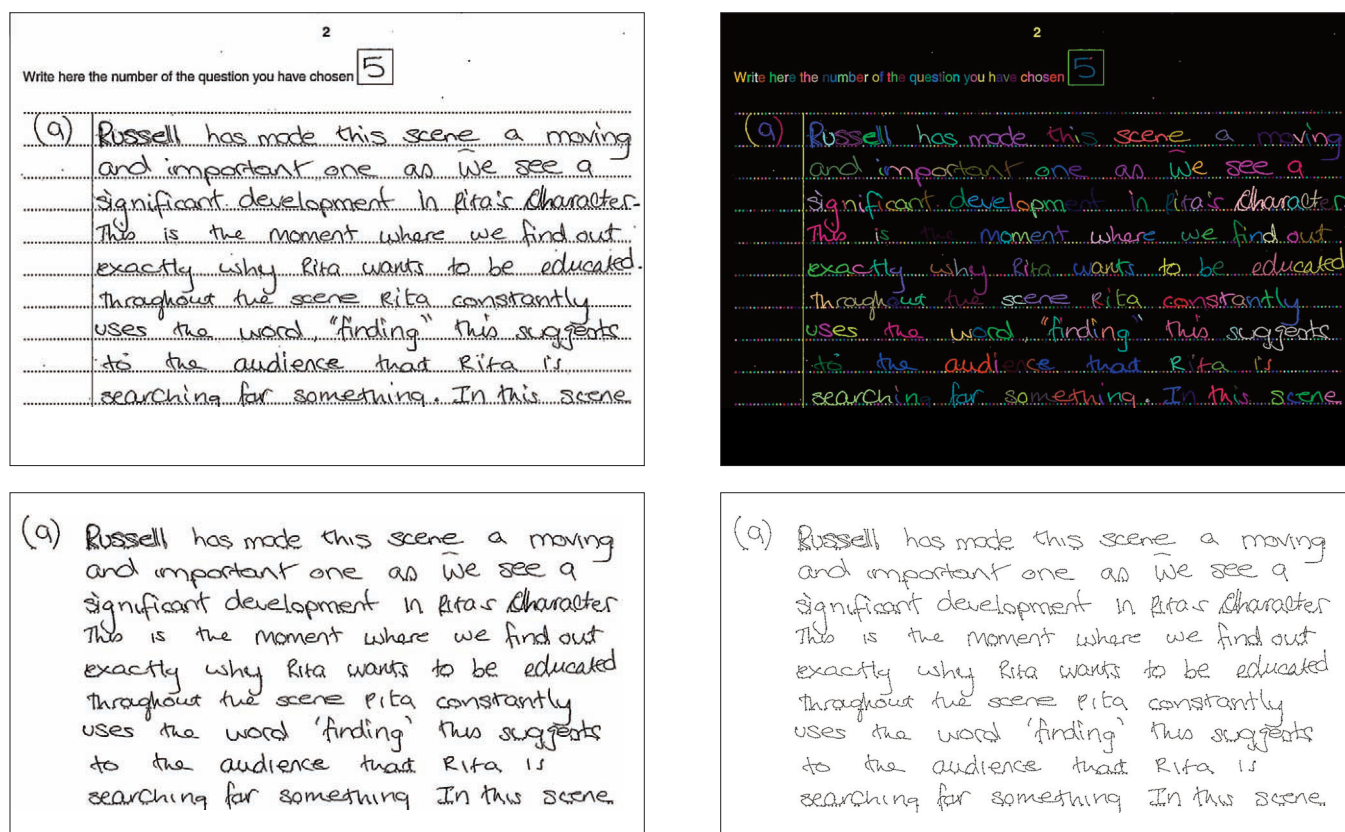


Figure 1: Steps in pre-processing images (Original image – top left; Segmented original image – top right; Cleaned image – bottom left; Thinned image – bottom right)

in the literature are fairly slow, requiring each pixel of an image to be considered in turn in relation to the surrounding pixels, and then either left alone or deleted as necessary. The decision for each pixel may then affect decisions for subsequent pixels. This means that the calculations need to be processed one at a time (at least within connected areas of an image).

As an alternative, an approximate but fairly fast approach was adopted. For the purposes of this method, the blurred density of each pixel was calculated by taking a weighted average of the pixels in the surrounding area (including the pixel itself) with more weight given to pixels that were nearby. Then, pixels of writing were only retained if the blurred density was greater than that of the pixels on either side in at least one direction. That is, the blurred density was either greater than both of:

- the pixel on the left and the pixel on the right, or
- the pixel below and the pixel above, or
- the pixel above and to the right and the pixel below and to the left, or
- the pixel above and to the left and the pixel below and to the right.

This method could be applied fairly quickly to each page and certainly helped address issues relating to the thickness of the pen. However, it could not be said to be a true 'thinning' method as the resulting image was often two or three pixels wide in certain areas rather than one. An example of why this occurred is given in Figure 2. This image represents a section of a letter 'p' from an image with the grey squares representing shaded pixels in the original image and the black squares those pixels that remain shaded after thinning. The numbers in the chart represent the blurred density at each point. As can be seen, in most places the image is reduced to be one pixel wide (in at least one direction). However, in the first two rows of the chart the image remains two pixels wide in all directions. This is because, whereas one pixel on the second row (with the value 0.634) has greater blurred density than that of the pixel on the left, this other pixel has greater density than the pixel below and to the right, as well as the pixel above and to the left. For this reason both pixels are retained and the shape is not completely thinned.

0.396	0.557	0.661	0.661	0.556	0.394	0.246	0.170
0.339	0.499	0.615	0.634	0.545	0.390	0.241	0.154
0.287	0.444	0.571	0.608	0.538	0.394	0.244	0.147
0.244	0.396	0.531	0.588	0.539	0.408	0.258	0.150
0.208	0.354	0.496	0.573	0.547	0.431	0.282	0.162
0.177	0.315	0.462	0.558	0.556	0.457	0.311	0.178
0.149	0.279	0.429	0.541	0.562	0.480	0.336	0.195
0.128	0.249	0.400	0.524	0.563	0.496	0.357	0.211

Figure 2: Example of the approximate thinning algorithm

Nonetheless, this algorithm can be applied fairly quickly, and, to a large extent, accounts for the thickness of the pen in any piece of writing. The bottom right-hand section of Figure 1 shows an example of a final image after thinning.

Having thinned the image, handwriting metrics are calculated based on each word. Ideally, we would rely on pupils having joined-up writing to identify words as any continuous sequence of marked pixels within the image. Sadly, as seen earlier, few candidates have completely joined-up writing. For example, in forming the letter 'f', most candidates will pause their writing to cross the letter rather than immediately joining on to the next one. Similar behaviour can also occasionally be found with letters such as 't', 'i' and 'j'. In addition, capital letters at the start of sentences are not usually joined to others in the same word. For this reason, a dilated version of the writing was created whereby, every time we find a marked pixel within the image, seven pixels to the left and seven to the right are also marked. For most candidates, applying this dilation step ensured that the majority of the letters within a word were joined together whereas separate words usually remained separated within the image. Note that, although this step is used to identify the location of words, the thinned version of the image produced in the previous step is also retained in order to calculate handwriting metrics. Figure 3 illustrates the thinning and word segmentation steps. The dilated image portions used to identify separate words are shown in different colours. Within each 'word' a blue line shows the thinned version of the handwriting.



Figure 3: Thinned writing within post-dilation connected areas

Metrics

Once pre-processing was complete, a number of metrics were calculated for both Unit A and Unit B, for each page submitted by each candidate.

The first two metrics were computed prior to the thinning step of pre-processing as this saved considerable time in computation and it was of interest to discover whether we could get a reasonable indicator of candidate identity at this stage. Specifically we calculated:

1. **Median pixels per line (PPL) prior to thinning:** Specifically, we restricted analysis to the rows of the matrix (representing the image) where the variance of the values in the row was greater than the median. This was done to ensure that rows that were either almost completely blank or (possibly) almost completely black (such as might occur at the margins due to scanning) were removed. The columns of the matrix were restricted in exactly the same way. Then, the proportion of pixels in each row that were black was calculated and the median of this value across all relevant rows was taken.

2. 80th percentile of pixels per line (PPL) prior to thinning:

Similar to Metric 1 but with the 80th percentile of the row density stored rather than the median. The idea behind using this metric was to ensure that density was being calculated within parts of the text where a full line of writing had been completed thus excluding shortened lines that might occur at the beginning or end of a paragraph.

A further five metrics were calculated after all the pre-processing steps (including thinning) had been completed.

3. **Rough word count:** This simply counted the number of separate joined sections of writing identified after the dilation step. For example, this would count the number of separate sections identified in Figure 3 (but across a full page). Technically, this metric is not entirely related to the handwriting style. However, it was useful for identifying which pages contained sufficient writing for meaningful analysis as well as being an interesting variable for analysis in its own right.

4. **Writing density within words (sometimes labelled within this article as 'word density'):** This metric calculated the percentage of pixels within the segments identified in the image that contained the thinned version of the writing after all holes within the segment had been filled in. For example, within the pink area identifying the first word ('Russell') in Figure 3, this metric calculates the percentage of the pixels that are covered by the thinned version of the writing. This metric is designed to distinguish writing that is small and tightly packed for writing that is large and loopy. The median of this value was taken across all of the words on the page.

5. **Standard deviation of writing density within words:** Similar to Metric 4 but, rather than focussing on the median density across words, this metric calculates the extent to which the writing density varies across the words. This metric was intended to capture the consistency of handwriting within the page.

6. **Area of words:** This metric separately calculated the number of pixels covered by each of the sections of the image identified as words (including dilation and filled holes). The median of this metric was taken across all of the words within the page. This metric was designed to measure the size of a candidate's writing.

7. **Perimeter of words:** Similar to Metric 6, and again designed to measure the size of writing, but calculated via the perimeter of the identified sections rather than the area. Once again the median value of the perimeter was taken across all of the words on a page.

The above metrics were calculated for each page in a candidate's response. In order to compare metrics between Unit A and Unit B it was necessary to reduce the data to one observation per candidate (rather than per page). This was done by removing any pages where the rough word count was below 10 and then taking the median value of each metric across the remaining pages. The only exception to this procedure was for word count where it was of interest to take the total word count across all pages (including those with less than 10 words) rather than the median word count per page.

Due to the obviously close relationship between area and perimeter, it was decided at the end of the calculations to combine the two to create one final metric:

8. **Shape:** This was defined as (median) word perimeter squared divided by (median) word area. This metric is similar to *circularity* (also known as the *isoperimetric quotient*) which is an existing measure of the shape of an object. In theory, this metric will assign higher values to writing that is low and wide than to writing that is tall and square.

The effectiveness of these different metrics is evaluated in the next section. However, from the metric descriptions, it is immediately obvious that these were not the only set of metrics that could have been chosen. For example, why focus on the median metric across words or lines of an image? Why not calculate metrics relating directly to the height and width of words? Should the density of pixels within a word be calculated within a dilated version of this same text, or should it be calculated within a box defined by the top, bottom, leftmost and rightmost pixels? The decisions that were made in regard to these questions were somewhat arbitrary and fairly strongly influenced by the availability of existing functions within the *EBImage* package to perform each task. Further research could explore the effect of different choices. However, as we will see later, the metrics performed relatively well and give a reasonable idea of what can be achieved using simple metrics.

Computing speed

Despite the relative simplicity of the described metrics, processing each page from each script was still relatively slow – taking around 7.5 seconds. Thus processing 6 pages for Unit A and 14 pages for Unit B took around 2.5 minutes for each candidate. Given that more than 26,000 candidates took both exams, more than 1,100 hours of computing time were required to compute all of the metrics for all candidates. In real terms this was reduced considerably by using multiple machines and splitting the processing across multiple cores on each machine. Nonetheless, processing these images was slow, requiring an entire weekend to calculate all of the necessary metrics for all candidates on both examination papers.

Results from the trials

Before beginning the analysis, any candidates with highly unusual handwriting metrics were removed from the data. This was done as the aim of the analysis was to identify candidates where the style of handwriting changed between occasions – not to simply identify scripts with very unusual handwriting or features. For this reason, any scripts where any of the described metrics were more than four standard deviations above the mean on either Unit A or Unit B were excluded². In particular, this process helped to remove atypical scripts where the response had been typed as well as other unusual cases, including one instance where the candidate had decided to draw a series of cartoon images (unrelated to the exam question) rather than write an essay. A total of 25,450 candidates were retained within the analysis.

Performance of metrics

The stability of each of the eight metrics between Unit A and Unit B is examined in Figure 4. As we can see, for each metric there is a clear positive correlation between the values calculated on each exam.

2. Since all of the metrics had a natural lower bound of zero, it was not necessary to exclude candidates with unusually low values.

The smallest correlation (0.46) is for the standard deviation of pixel density within words (Metric 5). However, all of the remaining seven metrics display a correlation greater than 0.8 between occasions, and four of them display correlations above 0.9. The highest correlation relates to median pixel density within words (Metric 4) which displays a correlation in excess of 0.95 between occasions. For comparison, the correlation between the marks awarded to candidates on Unit A and those achieved on Unit B was just 0.51. In other words, the metrics of handwriting developed in the previous section are far more stable between examinations than the performance of candidates. All three of the most successful metrics were calculated after thinning had been applied to the images. This suggests that this is a worthwhile step.

Correlations between the different measures indicated that, to a large extent, they provided separate pieces of information about candidates' writing style. As might be expected, given that both metrics relate to word size, perimeter and word area displayed a correlation in excess of 0.9. In addition, perhaps due to the way perimeter was used in its definition, perimeter and shape had a correlation of 0.76. Pixel density within words had a negative relationship with both word area (-0.66) and word perimeter (-0.56) – the slightly obvious point being that candidates with bigger writing will tend to leave more space within the words themselves. Aside from these obvious relationships, the correlations between the different metrics tended to be small, with the majority being below 0.2.

Initial analysis attempted to make use of all of the above metrics simultaneously in order to identify candidates with a large change in handwriting style³. However, manual inspection of script images from the 20 candidates showing the biggest overall change from Unit A to Unit B revealed some problems with this approach. Only eight of these scripts displayed clearly different handwriting between the two occasions. In some other cases the style of handwriting was inconsistent within examinations rather than between, and in other cases they appeared to have been identified as different for reasons other than a change of handwriting style. For example, in two cases the handwriting looked similar but it was likely that a major change in the type of pen used for writing led to a major change in values for Metrics 1 and 2 – underlining the importance of the thinning step. In another two cases, an extreme change in the length of the submission (i.e., the word count) appeared to be the main reason for the candidate being identified, rather than any obvious difference in the style of the handwriting.

As an alternative, a second, much simpler, approach was adopted. The best metric from Figure 1 (pixel density within words) was chosen. It should be noted that the mean absolute difference in this metric between occasions for any candidate was just 0.004. In contrast, the mean absolute difference between two randomly chosen candidates was

3. Linear discriminant analysis (https://en.wikipedia.org/wiki/Linear_discriminant_analysis) was used to combine the metrics.

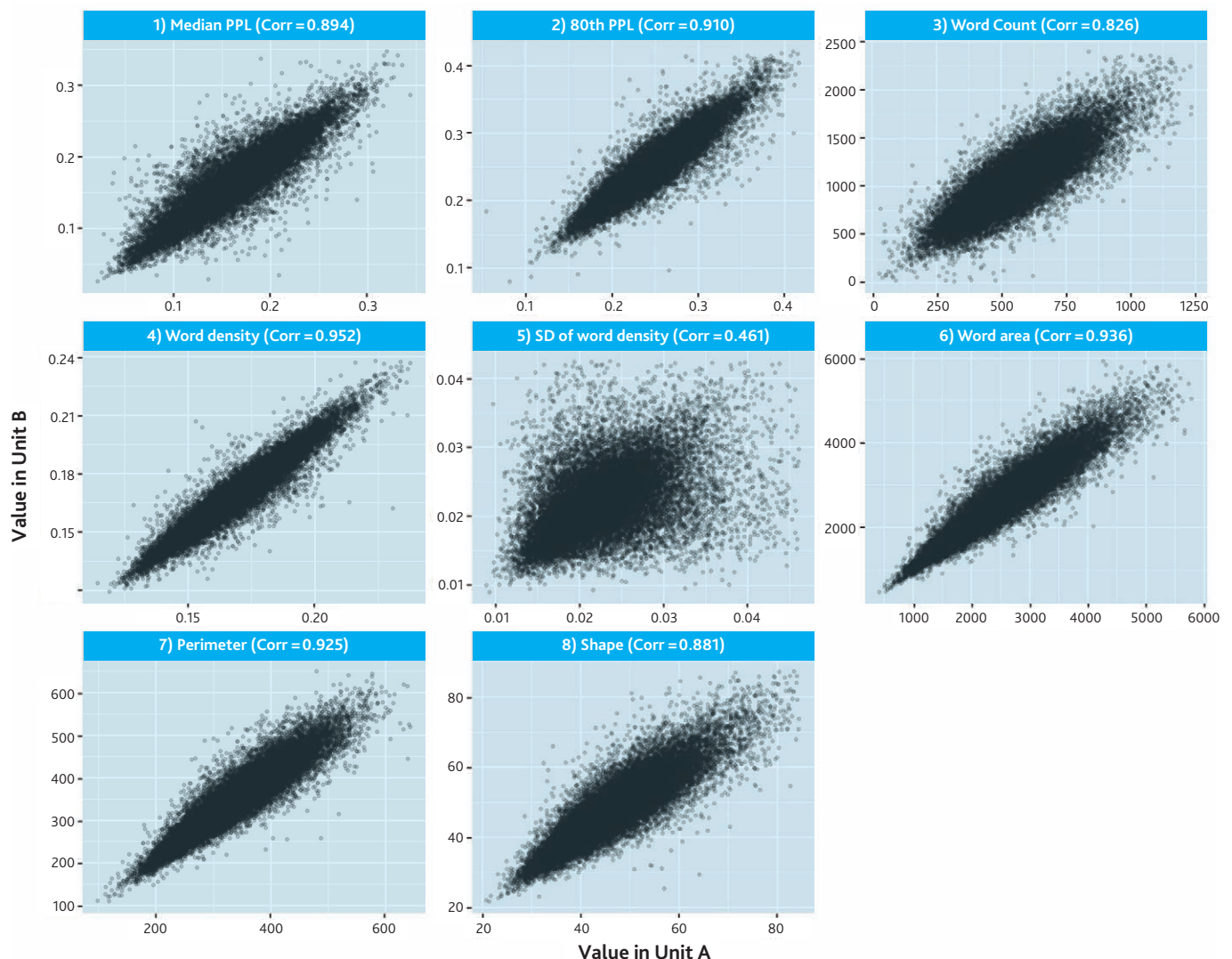


Figure 4: Relationship between each metric for Unit A and for Unit B

five times higher at 0.02. Only 213 candidates out of more than 25,000 with available data displayed such a large difference between occasions. This indicates that a focus upon this metric alone could yield interesting scripts for inspection.

Rather than examining all such instances, the 20 candidates' scripts showing the greatest change in this metric between Unit A and Unit B were inspected by eye. Fifteen of the scripts identified in this way had visibly different handwriting between the two occasions. Further details are given the next section.

Examples

Table 1 provides a list of the 20 candidates with the greatest changes in the chosen metric between occasions. The table begins with the candidate showing the greatest difference in pixel density within words between Unit A and Unit B and works through them in order, noting the qualitative impression of why the candidate has been identified, as well as the grades they achieved on each paper.

Table 1: Notes on 20 candidates with greatest difference in pixel density within words between Unit A and Unit B

Case No. (ranked starting from greatest difference)	Possible reason why identified	Grade on Unit A	Grade on Unit B
1	Visible difference in handwriting	C	B
2	Visible difference in handwriting	B	D
3	Visible difference in handwriting	A	D
4	Visible difference in handwriting	A	C
5	Visible difference in handwriting	B	D
6	Visible difference in handwriting	B	C
7	Visible difference in handwriting	B	C
8	Visible difference in handwriting	A	B
9	Visible difference in handwriting	A	A
10	Visible difference in handwriting	D	D
11	Visible difference in handwriting	B	E
12	Visible difference in handwriting	B	B
13	Visible difference in handwriting	U	E
14	Very little writing; Change of pen	D	U
15	Inconsistent handwriting	A*	A*
16	Inconsistent handwriting	B	D
17	Visible difference in handwriting	B	B
18	Visible difference in handwriting	B	A
19	Inconsistent handwriting	C	C
20	Not obvious why flagged	C	D

As noted above, in fifteen of the cases identified by this method the style of handwriting was visibly different between Unit A and Unit B. For example, Figure 5 compares part of the first page of writing on Unit A to part of the first page of writing on Unit B for the candidate with the largest change between occasions. As we can see, there is a marked difference in handwriting style. For Unit A, the handwriting is tidy, with curved characters and a uniform height. In contrasts in Unit B, the writing has a messy, uneven and angular style. The different styles shown in these small portions continued throughout the examination scripts. Nonetheless, having manually checked the names as well as the centre and candidate numbers entered on the front of both scripts, it is clear that both pieces of writing supposedly belong to the same candidate.

However, before leaping to the conclusion that one or other of these responses (or perhaps both) was provided by an imposter, there are

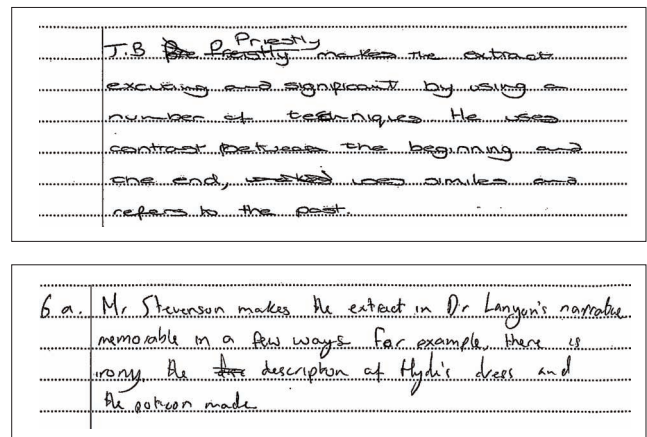


Figure 5: Portion of image of writing in Unit A (top) and Unit B (bottom) for the candidate with the greatest change in median pixel density per word

some other pieces of evidence to consider. To begin with it can be seen from Table 1 that, although the handwriting changed, the level of performance achieved was fairly similar on both examinations with a C grade awarded for Unit B and a B grade awarded for Unit A. This in itself suggests little motive for impersonation.

Secondly, examples of handwriting from other candidates examined as part of this research revealed cases potentially indicating that a single person might use very different handwriting styles in two examinations. An example (found in a separate analysis) is shown in Figure 6. Again, this example shows a marked change, from a large and looping style used in Unit A, to a small and neat style adopted in Unit B. However, the response to Unit B also displays another clear characteristic – the fairly large circles, almost like hollow umlauts, used to dot the 'i's. This same trait is also visible in Unit A. Given the unusual nature of this trait, it would appear at least possible that both sets of writing were produced by the same person. This suggests that we need to exercise some caution before concluding that a change in handwriting style indicates a change of author – an important fact when we consider Figure 5.

To emphasise this point further it is possible to find candidates where the style of handwriting changes even within the same examination.

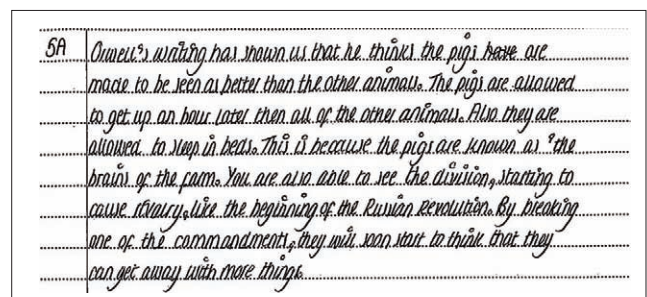
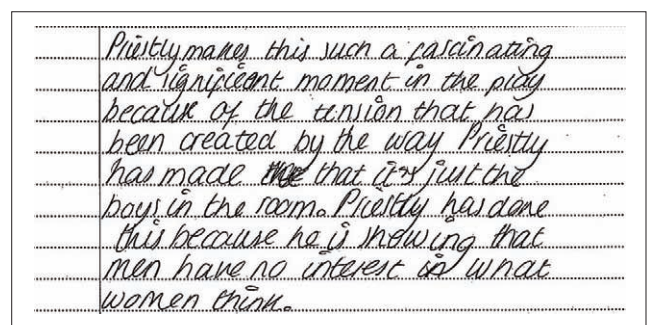


Figure 6: Example of a candidate with a consistent trait (the dots of the 'i's) but a different handwriting style (Unit A on top, Unit B on bottom)

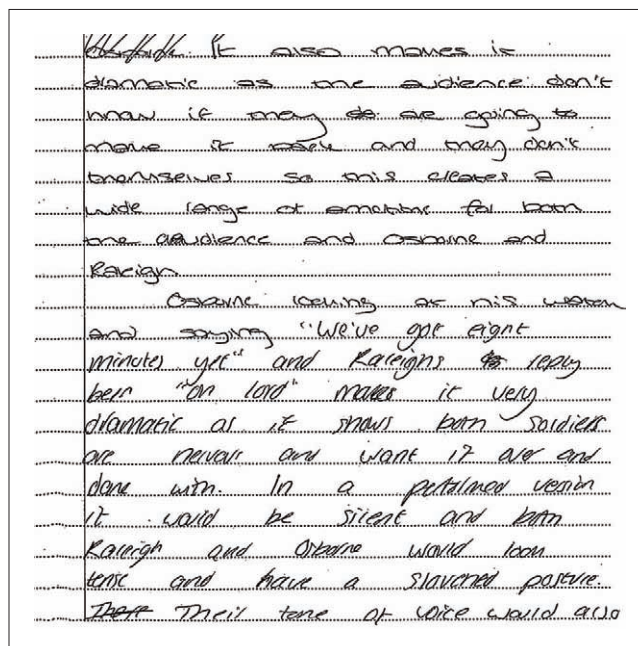


Figure 7: Example of candidate changing writing styles within the same page

This situation was evident in Cases 15, 16 and 19 in Table 1. Figure 7 shows a page from Case 19 that illustrates the issue most clearly. As we can see, the squat and curved lettering at the top of the figure gives way to taller and more angular writing at the bottom. We do not know what has caused the change although we might guess that time pressure or stress has led to a change of style. Indeed, one possibility that may be worth investigation is whether the writing at the end of this page was added by another author. However, this page does suggest a need for caution in interpreting the results. If handwriting can change even within a single page of writing, we cannot necessarily conclude that a change of handwriting style between Unit A and Unit B indicates any form of malpractice.

Returning to Table 1, it is notable that in nearly all cases the grade achieved on Unit A was similar to that awarded for Unit B. This fact, together with the examples discussed and the fact that this form of cheating is not widely reported in the UK, in any case suggest that something other than a change of the person taking the exam may explain most (and possibly all) of the cases identified in Table 1. It should be noted that other researchers in this area (see Dolega et al., 2008) have noted a "lack of stability of human handwriting" which fits with the results we see here.

Discussion

This article has proposed a number of metrics of handwriting style that are relatively easy to calculate. The majority of the suggested metrics displayed very high correlations between occasions suggesting that they may provide a reasonable indicator of whether the same candidate has indeed taken all of the relevant examinations leading to a qualification. The most effective metrics required thinning methods to be applied to the image as part of pre-processing indicating that, although

computationally burdensome, this is a worthwhile step. The most effective metric (median pixel density within words) displayed a correlation in excess of 0.95 between separate examination occasions.

Out of more than 25,000 candidates taking both of the exams being studied, the metrics allowed us to quickly find a number of examples where a candidate's handwriting style changed between occasions. However, the fact that we were able to identify cases where the handwriting style had changed, but other aspects of the writing indicated the author may have been the same, suggests that a change of handwriting style in itself is not proof of malpractice. The same applies in cases where the style of handwriting changed within an individual exam.

Of course, in the UK school context of the scripts analysed in this article, the use of an imposter for one exam but not another is rarely reported as an issue. As such, any cases where handwriting is identified to have changed are perhaps more likely to be explained by other factors than by the presence of an imposter in one or more exams, and the automated methods of checking handwriting styles we propose here are unlikely to be useful. However, in other contexts, where we are more suspicious that an imposter may be used for one or more exams, the methods we suggest here may be helpful as they provide a relatively quick means by which candidates displaying inconsistent handwriting between exams can be identified. Thus, in contexts where we are more worried about this form of cheating, this may provide an efficient means of identifying the candidates worthy of further scrutiny.

On a more general level, this research has begun to develop our expertise in processing images from the DSR to procure useful information about candidates' responses. For example, one by-product of this research has been to calculate a rough word count for candidates' essays – a potentially interesting variable for further research. Further work could build upon this basis to explore further automated methods of collating information from candidates' script images for use in research.

References

- Dolega, B., Agam, G., & Argamon, S. (2008). *Stroke frequency descriptors for handwriting-based writer identification*. *Proc. SPIE* 6815, Document Recognition and Retrieval XV, 681501 (January 28, 2008). doi:10.1117/12.767227
- Lam, L., Lee, S. W., & Suen, C. Y. (1992). Thinning methodologies – A comprehensive survey. *IEEE Transactions on pattern analysis and machine intelligence*, 14(9), 869–885. doi: 10.1109/34.161346
- Linear discriminant analysis. (n.d.). In *Wikipedia*. Retrieved January 12, 2017, from https://en.wikipedia.org/wiki/Linear_discriminant_analysis
- Otsu's method. (n.d.). In *Wikipedia*. Retrieved January 12, 2017, from https://en.wikipedia.org/wiki/Otsu's_method
- Pau, G., Fuchs, F., Sklyar, O., Boutros, M., & Huber, W. (2010). EBIImage – an R package for image processing with applications to cellular phenotypes. *Bioinformatics*, 26(7), pp.979–981. doi:10.1093/bioinformatics/btq046
- R Core Team. (2015). *R: A language and environment for statistical computing*. The R Foundation for Statistical Computing, Vienna, Austria. Available online at: <https://www.R-project.org/>.

Evaluating blended learning: Bringing the elements together

Jessica Bowyer Exeter University and Lucy Chambers Research Division (The study was completed when the first author was based in the Research Division)

Introduction

This article provides a brief introduction to 'blended learning', its benefits and factors to consider when implementing a blended learning programme. It then concentrates on how to evaluate a blended learning programme and describes a number of evaluation frameworks found in the literature. It concludes by introducing a new framework.

What is blended learning?

Blended learning is a mixture of online and face-to-face learning. In the literature, blended learning is also known as 'hybrid learning' or the 'flipped classroom'. Although there has been some debate about an exact definition, Boelens, Van Laer, De Wever, and Elen (2015) define blended learning as "learning that happens in an instructional context which is characterized by a deliberate combination of online and classroom-based interventions to instigate and support learning" (p.5). The online element should not solely be an addition to classroom-based teaching; rather, blended learning requires the effective integration of both virtual and face-to-face methods (Garrison & Kanuka, 2004). For example, a university lecturer placing some selected course materials, such as a course handbook, on a virtual learning environment (VLE) would not constitute a sufficient 'blend'.

Blended learning appears to be most commonly used in Higher Education (HE) or adult education. The majority of the research literature is within the United States, but there is a growing body of literature about the implementation of blended learning in HE courses within the UK. The prevalence of blended learning within HE means that there is very little research regarding the use or impact of blended learning for primary or secondary students. Given the different motivations and expectations of adult and secondary learners, the lack of representation of younger students in the literature may mean that it is difficult to draw conclusions about the potential impact of blended learning for students in compulsory education (Sparks, 2015). This should be borne in mind when reading the literature review that follows.

What are the benefits of blended learning?

Improved outcomes

There is some evidence that the introduction of blended learning can lead to improved course outcomes, in terms of higher student retention as well as increased pass rates. Studies by López-Pérez, Pérez-López, and Rodríguez-Ariza (2011) and Boyle, Bradley, Chalk, Jones, and Pickard (2003) found that the introduction of blended learning in HE courses improved retention and correlated with improvements in students' attainment. Additionally, Stockwell, Stockwell, Cennamo, and Jiang

(2015) found that blended learning courses improved attendance at face-to-face classes, in self-report measures of student satisfaction, and in examination performance.

For students from non-traditional backgrounds, the evidence suggests that blended learning can improve retention, although it may not improve attainment. Holley and Dobson (2008) introduced a blended learning programme during the first term to counteract low dropout rates at London Metropolitan University. These students were usually late entrants to HE and therefore drop out rates were high as students often struggled to make a successful transition to university study. Nevertheless, students who were introduced to a new blended learning environment during the first term were less likely than previous cohorts to leave their course before completion. Additionally, Hughes (2007) found that using blended learning to identify and support 'at-risk' students improved coursework submission rates, but had no significant effect on attainment.

Strategic use of classroom time

The improvement in course outcomes due to blended learning has been partially attributed to a more strategic use of classroom time. Garrison and Kanuka (2004) argue that blended learning is effective because it questions the traditional lecture-based teaching model, allowing classroom time to focus on more active and meaningful activities. This has been corroborated by Delialioğlu (2012), who found that problem-based, rather than lecture-based, blended learning had higher levels of student engagement. Online activities can be used to either reinforce learning undertaken in the classroom, or they can serve as a basic introduction to topics before they are covered in more depth in class.

If pre-reading material is placed on the VLE, then classroom time can focus on deeper analysis or discussion of the topics. This may also allow teachers or tutors to spend more individualised time with students in class, focusing on areas of particular difficulty. Aspden and Helm (2004) found that blended learning especially helped students who lived far away from campus use their time at university more effectively as they were able to engage with materials at home prior to attending class. Additionally, they found that students who were struggling with particular topics in class were able to participate and engage with online materials and thus grow in confidence. Alternatively, blended learning can take the form of the 'flipped classroom', where students engage with online lectures and textbook material at home, before participating in group discussion and problem-solving in class. This may have greater benefits for some subjects than others: Stockwell et al. (2015) reported that this mode of blended learning was particularly successful in Science education because it allowed teaching to shift away from the traditional textbook model, and students were thus able to engage with scientific concepts on a deeper level.

Nevertheless, this strategic use of classroom time relies on students' successful completion of online activities. Blended learning thus depends on students' capacity to adopt resilient learning strategies, as well as their self-motivation to complete the course. The literature suggests that whilst blended learning may be a valuable tool which enables students to work independently and develop their study skills, individuals will inevitably respond differently to this challenge. In Wivell and Day's (2015) study, students reported that self-motivation, self-reliance and the ability to work independently were essential to their success on the blended learning course. However, students who already struggled in the face-to-face delivery struggled to adapt to the demands of the blended programme. Moreover, Pérez and Riveros (2014) found that whilst a blended learning programme generally increased students' autonomy and responsibility for their learning, a common complaint from tutors was that some students did not engage with the online activities or complete the online assignments. Similar findings were reported by Chen and DeBoer (2015), who found that the most successful students were those who engaged more frequently with the online materials.

Consequently, as independent study skills and self-motivation are essential to students' success on blended learning programmes, it may be pertinent for providers to help students develop these skills by offering additional study skills sessions. Students' likely self-motivation should also be borne in mind when developing blended learning programmes. The age of students and the compulsory nature of online assignments may affect this. For example, HE students may be more self-motivated by being able to choose their course and will be used to a more independent style of learning, whilst secondary students may be less motivated to engage with the online elements as they are more familiar with a classroom or lecture-based model. Alternatively, making the online tasks compulsory, or contributory towards a student's final grade, may increase engagement and submission by offering higher extrinsic motivation.

Online discussion

A further potential benefit of blended learning is the additional opportunity for peer and tutor interaction through online discussion. Online discussion in blended learning can either be asynchronous (such as discussion boards) or synchronous (such as Instant Messaging). However, these potential benefits are perhaps the greatest source of contention in the literature, with studies differing in their findings regarding students' enjoyment and perceived utility of online discussion.

For groups who have few face-to-face classes together, online communication can facilitate a sense of community. Aspden and Helm (2004) found that online communication through a blended learning environment enabled students to make and maintain connections with other students and their learning institution even when off campus. However, they cautioned that blended learning could not counteract pre-existing negative relationships between teachers and students, and teachers need to engage with the online environment for blended learning programmes to be successful. Students in So and Brush's (2008) study were also more likely to report higher satisfaction with the blended learning programme if they perceived there to be high levels of collaborative learning online. Furthermore, Garrison and Kanuka (2004) argue that students' comments in asynchronous online discussion are more likely to be thoughtful and supported by evidence than face-to-face classroom discussion. Consequently, they argue that online

discussion in blended learning develops a community of inquiry, which in turn entails greater levels of cognitive learning and critical thinking.

Conversely, other studies have shown that, in practice, asynchronous communication is often neither enjoyed nor utilised by blended learning students. Taylor, Nelson, Delfino, and Han (2015) found that students reported that online discussion was the least useful element of their blended learning course. Similarly, Pye, Holt, Salzman, Bellucci, and Lombardi (2015) indicated that students were broadly ambivalent about the utility of online discussion. Only half of the students in their study reported having useful online discussions or using the online environment to work with others. Similar findings have been reported by Ginns and Ellis (2007) and So and Brush (2008).

Nevertheless, online communication in blended learning is not restricted to peer discussion and should also involve teachers and tutors. Although Reed's 2014 study of staff attitudes towards blended learning at a UK university found that they considered online discussion forums to be the least important elements of VLEs, blended learning offers the opportunity for teacher-student engagement outside of the classroom and enhanced feedback. The literature indicates that where students have been able to communicate with tutors online, they have found this useful (Hughes, 2007). Subsequently, for blended learning to be most useful, tutors should use the online environment to offer feedback on online work, and to assist with students' queries or problems. It is likely that tutors would need training in this area.

Implementing blended learning programmes

Implementing a blended learning programme requires coherent and co-ordinated planning to be successful. Garrison and Kanuka (2004) highlight the variety of policy issues that universities need to consider. These include strategic planning of financial, technical and human resources, course scheduling (e.g., if fewer face-to-face lectures will take place), and tutor and student support. These will all need careful consideration if universities and/or schools contemplate introducing blended learning elements.

Additionally, a recurrent theme in the literature is that for blended learning programmes to be successful, two things are essential:

1. Comprehensive teacher or tutor training
2. Ongoing evaluation.

Tutor or teacher training is especially critical in universities where teachers are responsible for curriculum and assessment design in addition to implementing blended learning. Reed (2014) found that HE staff identified a lack of staff support/training and a lack of skills as the biggest barriers to implementing blended learning programmes at their institution. Boyle et al. (2003) and Hughes (2007) suggest that their programmes would not have been successful without specialist training, cautioning that others wishing to introduce their own programmes should ensure that teaching staff are trained to deal with all aspects of blended learning.

Furthermore, the literature suggests that ongoing evaluation of blended learning programmes is essential when implementing new courses. Boyle et al. (2003) argue that implementation of blended learning should be reasonably conservative at first, to allow for appropriate tutor training and to allow students to adapt to new learning styles. Programmes should be adapted over a number of years

to meet specific student and tutor needs, and therefore ongoing evaluation is critical to the success of blended learning. Additionally, Pombo and Moreira (2012) suggest that ongoing evaluation, during task development rather than solely at the end of the programme, gives a more thorough and multi-faceted evaluation which in turn ensures the overall quality of the course. We discuss different methods for evaluating blended learning later.

Access to technology

The success of blended learning programmes inevitably relies on students' equitable access to technology. However, few studies have directly addressed whether access to home computers affects the perceived success of blended learning, or whether certain groups of students are disadvantaged. This is most likely because Internet and computer access in educational institutions has rapidly increased, and the vast majority of (if not all) schools and universities in the UK provide access to computers for students. Additionally, the most recent statistics indicate that 86% of UK households now have access to the Internet, up from 57% in 2006 (Office of National Statistics, 2015); although this leaves 14% of households without Internet access. Students and teenagers are the most prolific Internet users, the most recent large scale survey of Internet use found that 100% of university students and teenagers aged 14 and over had access to the Internet (Dutton, Blank, & Groselj, 2013). Additionally, 92% of students accessed the Internet on multiple devices, such as tablets and mobile phones. This indicates that, in the UK at least, the implementation of blended learning programmes is unlikely to be impeded by inequitable access to technology.

Evaluating blended learning

As Pombo and Moreira (2012) indicate, there are four elements that need to be taken into consideration when evaluating blended learning programmes:

1. What is the purpose of evaluation?
To improve student engagement, resources, or overall course quality?
2. Who should be involved?
Lecturers, students, course leaders?
3. How and when should evaluation take place?
Methods of data collection; during the course or at the end?
4. What should be evaluated?
Teaching, learning, course outcomes, resources, quality of assessment?

The literature offers several methods of evaluating blended learning programmes. These differ in their methods (e.g., which data they use), which aspects of blended learning are focussed on (e.g., technology, course content), whose viewpoints are considered (e.g., students', teachers', administrators') and the criteria used to make judgements about the success of particular programmes. Generally, evaluation criteria include a combination of data about course outcomes (attendance, retention and students' marks) and measures of student satisfaction and student engagement.

Measuring course outcomes

A number of measures can be used to evaluate course outcomes: these include grades and marks, activity, attendance, and drop out rates. Measurement can be enhanced and made easier by use of the blended

learning system as student activity and results can be captured by the system. Using outcome measures alone may not give the full picture due to the effect of motivation: statistical measures do not capture students' attitudes towards learning and the role of the blended learning system in facilitating this. Consequently, Liu, Bridgeman, and Adler (2012) note that "accountability initiatives involving outcomes assessment should also take into account the effect of motivation when making decisions about an institution's instructional effectiveness" (p.360).

Measuring learner satisfaction

An important course outcome that cannot be measured through attendance and assessment data is learner satisfaction. Whilst a researcher or teacher might consider a course to be successful if students meet or exceed expectations in assessment, learner satisfaction is important because it accounts for students' personal experiences of the course. This is becoming particularly pertinent in HE in the UK, where the National Student Survey (NSS) is a key measure of perceived quality from students' perspectives. The NSS covers teaching, assessment, support, organisation, learning resources, personal development and overall satisfaction (IpSOS MORI & HEFCE, 2016). These results are made available to prospective students through an independent website, *Unistats*, and headline measures of overall satisfaction are often promoted on universities' own websites and prospectuses. Additionally, learner satisfaction, as measured through the NSS, will become more important in the future, as the government introduces the Teaching Excellence Framework (TEF). The TEF is intended to provide a measure of teaching quality at all UK universities and will be used to justify institutional fee increases (Department for Business, Innovation and Skills, 2016).

Common measures of learner satisfaction in blended learning courses use self-report questionnaires to investigate how satisfied students were with the course overall, the perceived quality of teaching, and, in particular, their experience of the blended learning environment. The specific items vary depending on the purpose of the evaluation and the researcher's personal perspective, but there tend to be similarities between studies. For example, Shee and Wang (2008) and Wang (2003) explicitly focus on students' experiences in an online learning environment and subsequently focus on the learning community, the learner interface, the course content, and the personalisation of the online environment. However, whilst Sun, Tsai, Finger, Chen, and Yeh (2008) name their elements of learner satisfaction as the learner, instructor, course, technology, design and environment dimensions, they investigate similar factors to Shee and Wang (2008), such as relationships between peers and teachers, perceived ease of use of technology, and course flexibility. Consequently, for measures of learner satisfaction to be appropriate within a blended learning environment, they should investigate students' perceptions of the ease of use of the technology and online content, in addition to teaching quality and overall experiences of the course.

Measuring student engagement

Measuring student engagement allows a more complex analysis of students' experiences and learning than simply investigating course outcomes. Engagement is "more than involvement or participation – it requires feelings and sense-making as well as activity" (Trowler, 2010, p.7). Understanding engagement has become particularly important in

the HE sector, as universities now operate in a more competitive marketplace. Consequently, measuring and improving student engagement can be an institutional advantage when attracting and retaining students (Trowler, 2010). Fredricks, Blumenfeld, and Paris (2004) identified three elements of student engagement: behavioural, emotional and cognitive. These are now widely accepted, although there remains some debate about how these can be most accurately defined and measured. Generally, they can be defined as:

1. **Behavioural:** relating to students' actions. For example, class attendance, submission of work, contribution to class discussion, or participation in school-related activities (e.g., extra-curricular sports or school governance).
2. **Emotional:** relating to students' affective reactions in relation to their learning. For example, an emotionally engaged student might report that they were interested in their course and that they enjoyed learning.
3. **Cognitive:** relating to students' psychological investment in their learning. For example, the desire to go beyond the requirements of the class and the adoption of metacognitive learning strategies.

It is important to note that engagement does not always have to be positive: a student could be negatively engaged if they report dislike or anxiety towards their learning. Trowler (2010) identifies positive and negative elements of all three definitions (see Table 1).

Table 1: Examples of positive and negative engagement (Trowler, 2010)

Reproduced courtesy of the author.

	<i>Positive engagement</i>	<i>Non-engagement</i>	<i>Negative engagement</i>
Behavioural	Attends lectures, participates with enthusiasm	Skips lectures without excuse	Boycotts, pickets or disrupts lectures
Emotional	Interest	Boredom	Rejection
Cognitive	Meets or exceeds assignment requirements	Assignments late, rushed or absent	Redefines parameters for assignments

Behavioural engagement has typically been investigated through student or teacher questionnaires, or classroom observations. It is also probably the easiest element of engagement to measure, as quantitative measures of attendance and submission of work can be used. Blended learning programmes can provide particularly rich data as it is possible to collect information about students' use of the online environment, including the frequency and duration of use. This may provide more objective data than self-report questionnaires.

Emotional and cognitive engagement are usually measured through questionnaires and interviews. Measuring emotional engagement is largely self-explanatory: students are asked about their feelings towards various aspects of their learning and classroom experience. Conversely, cognitive engagement is particularly difficult to measure, predominantly due to the inherent difficulty of assessing cognition. Consequently, measures of cognitive engagement predominantly rely on questionnaire items that aim to assess whether students are using deep or surface-learning strategies (Fredricks et al., 2004).

Existing evaluation frameworks

The majority of the literature evaluating blended learning has used a combination of author-designed questionnaires and course outcomes data. Students' opinions and experiences are often prioritised over those of teaching staff, and researchers have more often used questionnaires than interviews and focus groups. Several authors have created instruments for this purpose. These are typically either student questionnaires or rubric-based frameworks for evaluation by a researcher. Due to the diversity of methods and evaluation frameworks utilised in the literature, there is no one particular instrument that is seen to be the most effective for evaluating blended learning. We discuss some selected instruments and frameworks later in the article.

Web-Based Learning Environment Instrument (WEBLEI)

The WEBLEI is essentially a questionnaire investigating students' perceptions and experiences of online learning environments. It is divided into four areas or 'scales': the first three are based on categories in Tobin's (1998) qualitative evaluation of an online learning programme and the fourth focuses on information structure and design (Chang, 1999). The WEBLEI scales are: *Emancipatory activities* (looking at convenience, efficiency and autonomy); *Co-participatory activities* (looking at flexibility, reflection, quality, interaction, collaboration and feedback); *Qualia* (looking at success, confidence, accomplishments and interest); and *Information structure and design* (looking at how well the course and learning materials are structured and designed), (Chang, 1999). The scales are scored using a five-point Likert scale (Chang & Fisher, 2003). Some studies have used an additional survey with open-ended questions for a more in-depth analysis (see Chandra & Fisher, 2009).

Hexagonal E-Learning Assessment Model (HELAM)

HELAM is a conceptual multidimensional model for evaluating learning management systems in terms of perceived learner satisfaction (Ozkan & Koseler, 2009). It contains six dimensions (see Figure 1) assessed via a questionnaire. The instrument has been validated and all six dimensions were found to be important. The authors note the model is based on student perceptions only and does not consider the perceptions of other stakeholders such as teachers, system developers and administrators.

E-Learning framework

The E-Learning framework contains eight dimensions which can be used to "provide guidance in the design, development, delivery and evaluation of open and distributed learning environments." (Khan, n.d., para. 4). The dimensions are systemically interconnected to support learning (Figure 2) and are expanded in Table 2. The framework has been used to evaluate blended learning (e.g., Deegan, Wims & Petiti, 2015, and Gomes & Panchoo, 2015). The framework does not appear to contain any instruments for evaluation but provides a guiding structure with which to construct an evaluation.

Technology Acceptance Model (TAM)

A number of studies have focused solely on the technology aspects of blended learning and how they affect user satisfaction and course retention (Ma, Chao, & Cheng, 2013; Padilla-Meléndez, Del Aguila-Obra, & Garrido-Moreno, 2013). The Technology Acceptance Model (TAM)

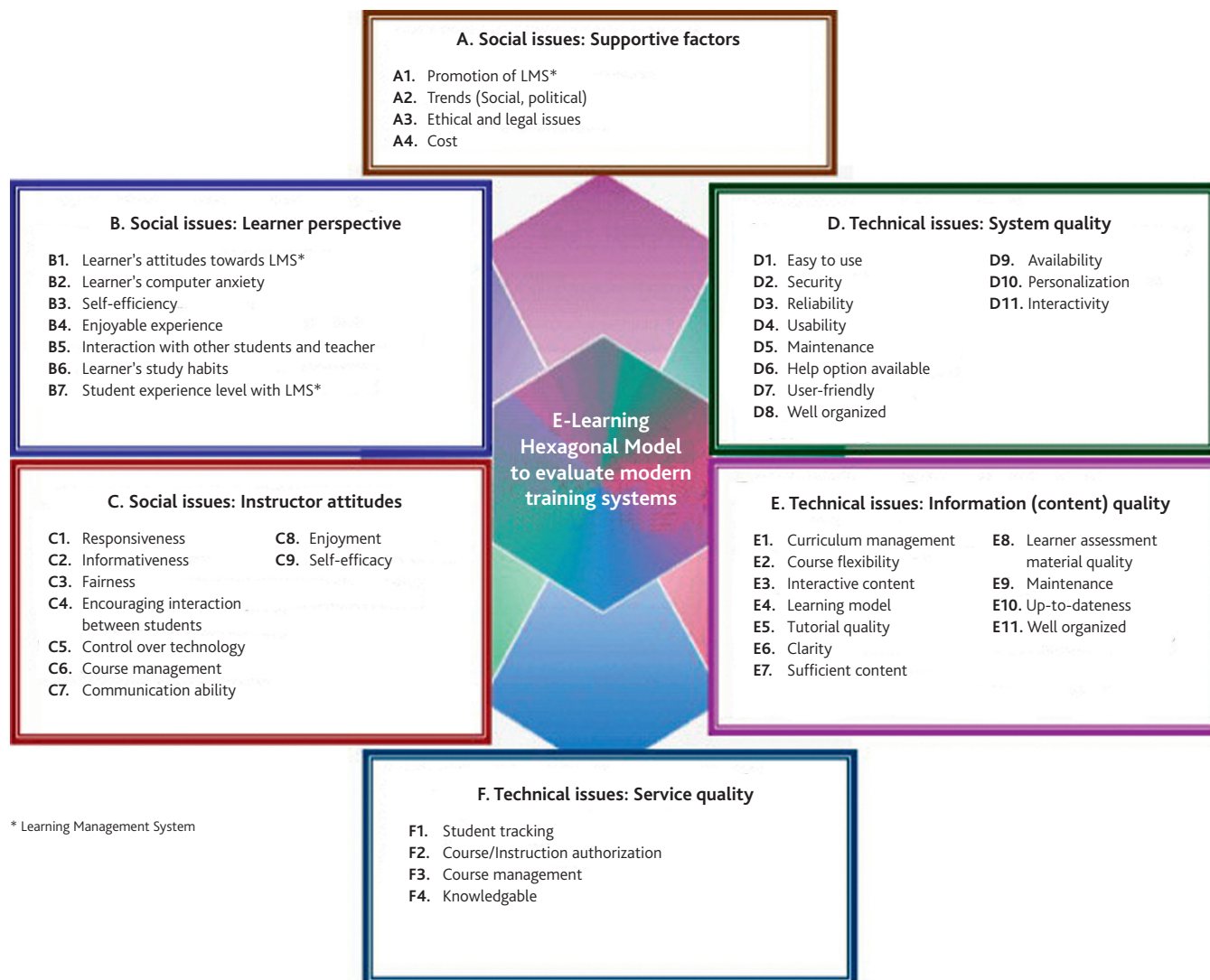


Figure 1: HELAM (Hexagonal E-Learning Assessment Model) (Ozkan & Koseler, 2009) Reproduced with permission from Elsevier.



Figure 2: E-Learning framework (Khan, n.d.) Reproduced with permission of the author under the Fair Use Policy.

Table 2: E-Learning framework

Adapted from Khan (n.d.) under the Fair Use Policy.

Dimension	Category
1. Pedagogical	Content analysis, audience analysis, goal analysis, media analysis, design approach, organization and methods and strategies of e-learning environments.
2. Technological	Infrastructure planning, hardware and software.
3. Interface design	Page and site design, content design, navigation, and usability testing.
4. Evaluation	Assessment of learners and evaluation of the instruction and learning environment.
5. Management	Maintenance of learning environment and distribution of information.
6. Resource support	Online support and resources required to foster meaningful learning environments.
7. Ethical	Social and political influence, cultural diversity, bias, geographical diversity, learner diversity, information accessibility, etiquette, and the legal issues.
8. Institutional	Administrative affairs, academic affairs and student services related to e-learning.

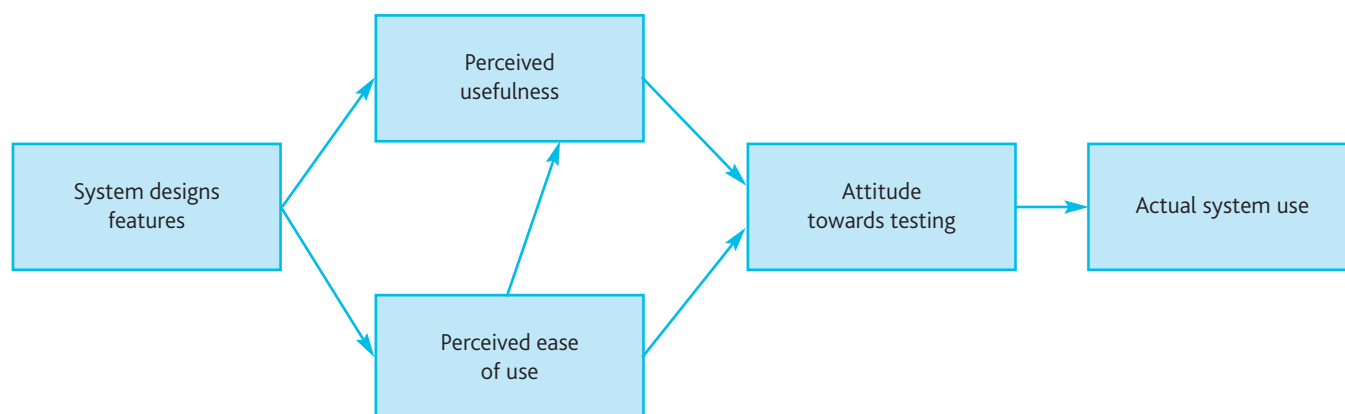


Figure 3: Technology Acceptance Model (Davis, 1993) *Reproduced with permission. Copyright, INFORMS®, <http://www.informs.org>*

(Davis, Bagozzi, & Warshaw, 1989) "specifies the causal relationships between system design features, perceived usefulness, perceived ease of use, attitude toward using, and actual usage behavior" (Davis, 1993, p.475). This is depicted in Figure 3.

Perceived usefulness (the degree to which a person believes that using a particular system would enhance their performance) and perceived ease of use (the degree to which a person believes that using a particular system would be free from effort) are two of the main predictors of system use (Padilla-Melendez et al., 2013). Caution should be taken if evaluating a blended learning programme solely on the basis of technological aspects as there are many other facets that influence programme effectiveness.

Rubric-based frameworks

Several researchers have created standards or rubric-based frameworks for evaluating blended learning environments. These are judgement-based and tend to be comprehensive in scope. Smythe (2011) argues that rubrics should be used as they cover a broad range of factors, such as instructional design and the use of technology, in addition to students' experiences of the programme. They are also beneficial as they provide a quick and efficient way for course designers to evaluate their programmes. Table 3 illustrates the factors measured by a selection of rubric frameworks and Figure 4 shows an example rubric (from the *Rubric for Online Instruction*, CSU, Chico, Copyright 2003/Revised 2009).

However, a key problem with rubrics is that they are inherently subjective due to their reliance on judgements. Although the example in Figure 4 uses criteria such as 'limited', 'adequate' and 'extensive', these terms are open to interpretation. Additionally, designers of rubrics do not provide advice about which data should be used to make judgements or how such information should be collected. This is especially pertinent when incorporating measures of student engagement in rubrics: should course designers conduct engagement questionnaires in order to provide a more accurate judgement? Consequently, whilst rubrics can provide a quick and broad overview of a blended learning programme, they lack the depth with which to fully evaluate the delivery of these programmes.

A new framework

As seen in the previous section, there are numerous frameworks and instruments for evaluating blended learning, although no particular one seems to be favoured in the literature. This is partly due to the diversity of reasons for evaluating blended learning systems, as well as the

Table 3: Dimensions measured by a selection of rubric frameworks

Author	Dimensions
California State University (2009)	Learner support and resources; online organisation and design; instructional design and delivery; assessment and evaluation of student learning; innovative teaching with technology; faculty use of student feedback.
Illinois Online Network (2008)	Instructional design, communication, interaction and collaboration; student evaluation and assessment; learner support and resources.
Maryland Online (2009)	Course overview and introduction; learning objectives; assessment and measurement; resources and materials; learner engagement; course technology, learner support; accessibility.
Mirriahi, Alonzo and Fox (2015)	Resources; activities; support; assessment.
Smythe (2011)	Student support and resources; course organisation; instructional design – learning objectives; instructional design – student engagement; assessment and evaluation of learning; use of technology.
The Sloan Consortium (2011)	Institutional support; technology support; course development and instructional design; course structure; teaching and learning; faculty support; student support; evaluation and assessment.

many intended audiences and perspectives for these evaluations. For example, some frameworks focus on technology over pedagogy, most focus on the student perspective rather than that of teachers or administrators, and some frameworks rely only on course outcome measures. Purpose also varies: some evaluations are designed for accountability, some for improvement, and others for marketing. However, we feel that it is important that any framework encompasses all aspects of the blended learning situation so that the interconnectedness is not lost. This approach still enables individual evaluations to focus on specific elements of a blended learning programme, but allows the researchers to see where these elements are situated within the wider context of blended learning, subsequently making it easier to identify omissions and acknowledge limitations. Additionally, we believe that a coherent overall framework permits researchers and evaluators to easily identify the relationships between different aspects of blended learning systems, such as between the institutional context and the support tutors are given when designing and implementing a blended learning programme.

One way to conceptualise this is to categorise a framework into

Category 1	Baseline	Effective	Exemplary
Learner Support and Resources	A. Course containing limited information for online learner support and links to campus resources.	A. Course contains adequate information for online learner support and links to campus resources.	A. Course contains extensive information about being an online learner and links to campus resources.
	B. Course provides limited course-specific resources, limited contact information for instructor, department, and/or program.	B. Course provides adequate course-specific resources, some contact information for instructor, department, and program.	B. Course provides a variety of course-specific resources, contact information for instructor, department, and program.
	C. Course offers limited resources supporting course content and different learning abilities.	C. Course offers access to adequate resources supporting course content and different learning abilities.	C. Course offers access to a wide range of resources supporting course content and different learning abilities.

Figure 4: Example rubric for evaluating online learning environments (California State University, 2009)

Reproduced under the Creative Commons Attribution 3.0 United States License, <https://creativecommons.org/licenses/by/3.0/us/>

spheres of concentric influence¹ so that any evaluation can focus on a particular perspective but acknowledge the influence of other elements of the framework (see Figure 5). Three spheres of influence have been identified, each containing a number of elements. The outer sphere is *situation*: this encompasses the wider context as well as institutional elements. The mid-sphere is *course organisation*: this contains design and planning, content, technology and assessment. The inner sphere is *individual perspectives*: this focuses on the learner and teacher elements but also contains the crucial features of communication, interaction and collaboration which operate at this level.² These described spheres can

be thought of as the independent variables: the inputs and processes that form the facets of the blended learning programme. There is also the core of the sphere: this contains the *outcomes*, namely learner satisfaction, student engagement and course outcomes. These can be considered the dependent variables. These spheres and elements are detailed in Table 4, which also includes suggestions for measurement. An additional feature that runs throughout the framework is support. This is vital for a successful blended learning programme and should be conceptualised as influencing elements of each sphere, as well as the relationships between spheres. There is an inevitable interaction between institutional support, tutor support and student and tutor experiences. For example, a learner can receive financial support to take a course (context), careers support (institution), special needs support (design and planning), tailored learning (content), IT support (technology), formative tests (assessment), peer feedback (learner), and feedback on learning (teacher). Consequently, although support does not constitute its own element or sphere within the framework outlined, elements of support should be investigated in all three spheres.

The framework outlined here was developed by looking at many of the existing frameworks for evaluating blended and e-learning, listing all the constructs encapsulated by them and adding others that we considered to be missing. These were then grouped into spheres at the situation, course and individual level to develop what we consider to be a coherent overall framework. We believe this framework can be used beyond blended learning and can be applied to other technology-based learning situations.

1. 'Spheres of influence' is a term traditionally used in international relations. Its use here has no political basis.
2. This framework has parallels with a context-based model for investigating impact in educational systems used by Cambridge English Language Assessment (Saville, 2010). The model stresses the dynamic interplay between the multiple macro (e.g., country, region, community and school) and micro (e.g., learner, teacher and class) contexts.

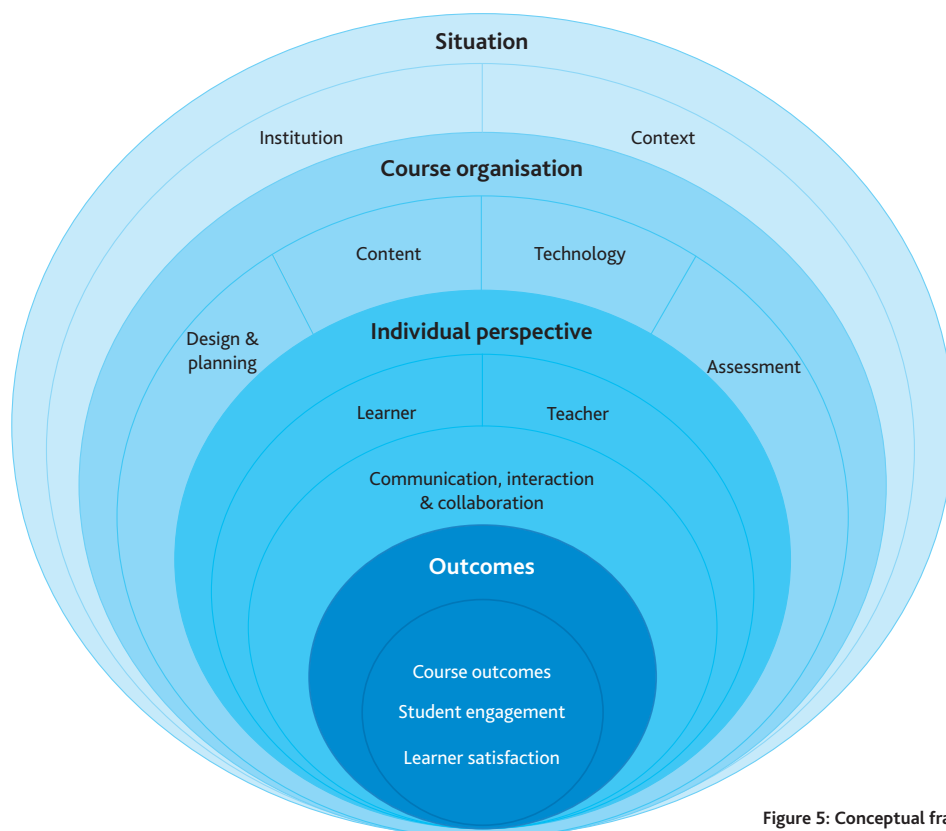


Figure 5: Conceptual framework for evaluating blended learning

Table 4: Framework for evaluating blended learning

<i>Level</i>	<i>Variable</i>	<i>Elements</i>	<i>Measurement</i>
Situation	Context	Socio-economic Ethical Legal Cost Accessibility Cultural Geographical Support	Can be investigated by independent evaluation based on full knowledge of the programme's context, but more likely through interview with, or questionnaire for, course administrators and/or teachers.
	Institution	Support Administration	Can be measured through self-report questionnaires, interviews or focus groups with course administrators and/or teachers.
Course	Design and planning	Curriculum management Organisation of teaching (the 'blend') Flexibility Support	Can be investigated by independent evaluation based on full knowledge of the programme's context, using course materials, but more likely through interview with, or questionnaire for, course administrators and/or teachers.
	Content (online and in class)	Relevance and scope Quality Breadth of content Breadth of methods of presentation and activities Validity Accuracy and balance Interactivity Accessibility Organisation Currency (up-to-dateness) Support	Can be measured through independent evaluation of the blended learning platform and course materials (in relation to curriculum or specification documents) or self-report questionnaires (from students). Existing elements from the latter could be taken from: • HELAM: Technical issues – information (content) quality • WEBLEI: Information structure and design activities.
	Assessment	Diversity Fit/relevance Support	Can be measured through independent evaluation of the blended learning platform and course materials (in relation to curriculum or specification documents) or self-report questionnaires (from students).
	Technology	Interface design Ease of use Security Reliability Usability Maintenance Accessibility Organisation Availability Personalisation Interactivity Currency (up-to-dateness) Support	Can be measured through independent evaluation of the platform or self-report questionnaires. Elements of the latter could be taken from: • HELAM: Technical issues – system quality • WEBLEI: Information structure and design activities • WEBLEI: Qualia • Online engagement scale (Krause & Coates, 2008) • The Technology Acceptance Model (Davis, 1993) can be used to explore the influence of technology.
Individual	Teachers	Attitude towards computers and technology Attitude towards learners Technological experience Teaching experience Subject knowledge Response time* Feedback* Support Provision of information	Can be measured through questionnaires, interviews and focus groups. Response time and feedback can be investigated using online platform data. There are few published instruments focussing on teacher perspectives. • Reed (2014): Learners' attitudes to technology in education.
	Learners	Attitude towards computers/technology Attitude towards learning Attitude towards teaching staff Motivation to take the course Study habits Technological experience Prior knowledge & learning experience Convenience Autonomy Perceived usefulness Perceived enjoyment Peer interaction/support* Group working and collaboration*	Can be measured through self-report questionnaires, interviews and focus groups. Existing elements could be taken from: • Peer-engagement scale (Krause & Coates, 2008) • Student-staff engagement scale (Krause & Coates, 2008) • WEBLEI: Co-participatory activities • WEBLEI: Emancipatory activities • HELAM: Learner's perspectives • HELAM: Instructor attitudes • Sun et al. (2008): Learners' attitudes to technology.

Table 4: Framework for evaluating blended learning (continued)

Level	Variable	Elements	Measurement
Outcomes	Learner satisfaction	With course (overall) With learning With teaching Utility of course for future plans/education	Can be measured through self-report questionnaires. Existing elements could be taken from: • Sun et al. (2008): Perceived learner satisfaction • NSS: Contribution of course to knowledge, skills and development.
	Student engagement	Psychological and cognitive engagement Behavioural engagement Emotional engagement	Can be measured through self-report questionnaires. Behavioural engagement can be investigated using online platform data. Existing elements could be taken from: • Academic engagement scale (Krause & Coates, 2008): psychological/cognitive engagement • Intellectual engagement scale (Krause & Coates, 2008): psychological/cognitive engagement • NSS: Elements of behavioural and cognitive engagement.
	Course outcomes	Grades and marks Online activity Attendance Drop out rates	Can be measured using the online platform data and teacher reports.

*Note: These elements entail the communication, interaction and collaboration aspect of the framework.

References

- Aspden, L., & Helm, P. (2004). Making the Connection in a Blended Learning Environment. *Educational Media International*, 41(3), 245–252. Retrieved from <http://www.tandfonline.com/doi/abs/10.1080/09523980410001680851>
- Boelens, R., Van Laer, S., De Wever, B., & Elen, J. (2015). *Blended learning in adult education: towards a definition of blended learning*. Retrieved from <https://biblio.ugent.be/publication/6905076>
- Boyle, T., Bradley, C., Chalk, P., Jones, R., & Pickard, P. (2003). Using blended learning to improve student success rates in learning to program. *Journal of Educational Media*, 28(2–3), 165–178. Retrieved from <http://www.tandfonline.com/doi/abs/10.1080/1358165032000153160>
- California State University (CSU). (2009). Chico. *Rubric for Online Instruction*. Retrieved from <http://www.csuchico.edu/eoi/documents/rubricpdf>
- Chandra, V., & Fisher, D. L. (2009). Students' perceptions of a blended web-based learning environment. *Learning Environments Research*, 12(1), 31–44. Retrieved from <http://link.springer.com/article/10.1007/s10984-008-9051-6>
- Chang, V. (1999). Evaluating the effectiveness of online learning using a new web based learning instrument. *Proceedings Western Australian Institute for Educational Research Forum 1999*. Retrieved from <http://www.waier.org.au/forums/1999/chang.html>
- Chang, V., & Fisher, D. L. (2003). The validation and application of a new learning environment instrument for online learning in higher education. In M. S. Khine & D. L. Fisher (Eds.), *Technology-rich learning environments: A future perspective* (pp.1–20). Singapore: World Scientific Publishing Co. Pte. Ltd.
- Chen, X., & DeBoer, J. (2015). *Checkable answers: Understanding student behaviors with instant feedback in a blended learning class*. 2015. IEEE Frontiers in Education Conference (FIE), 00(undefined), 1–5. Retrieved from <http://doi.ieeecomputersociety.org/10.1109/FIE.2015.7344045>
- Davis, F. D. (1993). User acceptance of information technology: system characteristics, user perceptions and behavioral impacts. *International Journal of Man-Machine Studies*, 38(3), 475–487. Available online from doi.org/10.1006/imms.1993.1022
- Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User Acceptance of Computer Technology: A Comparison of Two Theoretical Models. *Management Science*, 35(8), 982–1003. Retrieved from http://www.jstor.org/stable/2632151?seq=1#page_scan_tab_contents
- Deegan, D., Wims, P., & Pettit, T. (2015). The Potential of Blended Learning in Agricultural Education of Ireland. *International Journal of Agricultural Science, Research and Technology in Extension and Education Systems*, 5(1), 53–64. Retrieved from http://ijasrt.iau-shoushtar.ac.ir/article_520557_8f0f5708e327c982c75fc60764636ed0.pdf
- Delialioğlu, Ö. (2012). Student Engagement in Blended Learning Environments with Lecture-Based and Problem-based Instructional Approaches. *Journal of Educational Technology & Society*, 15(3), 310–322. Retrieved from <http://www.jstor.org/stable/jeductechsoci.15.3.310>
- Department for Business, Innovation and Skills. (2016). *The Teaching Excellence Framework: Assessing quality in Higher Education*. London, UK: The Stationery Office (TSO). Retrieved from <http://www.publications.parliament.uk/pa/cm201516/cmselect/cmbis/572/572.pdf>
- Dutton, W., Blank, G., & Groselj, D. (2013). Cultures of the Internet: The Internet in Britain. *Oxford Internet Survey 2013*. Oxford Internet Institute, University of Oxford. Retrieved from <http://oxis.oii.ox.ac.uk/wp-content/uploads/2014/11/OxIS-2013.pdf>
- Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School Engagement: Potential of the Concept, State of the Evidence. *Review of Educational Research*, 74(1), 59–109. Retrieved from <http://rer.sagepub.com/content/74/1/59.short>
- Garrison, D. R., & Kanuka, H. (2004). Blended learning: Uncovering its transformative potential in higher education. *The Internet and Higher Education*, 7(2), 95–105. Available online from doi.org/10.1016/j.iheduc.2004.02.001
- Ginns, P., & Ellis, R. (2007). Quality in blended learning: Exploring the relationships between on-line and face-to-face teaching and learning. *The Internet and Higher Education*, 10(1), 53–64. Available online from doi.org/10.1016/j.iheduc.2006.10.003
- Gomes, T., & Panchoo, S. (2015). *Teaching Climate Change Through Blended Learning: A case study in a Private Secondary School in Mauritius*. Paper presented at the 2015 International Conference on Computing, Communication and Security (ICCCS), (pp.1–5). IEEE. Retrieved on 14 April 2016 from <http://ieeexplore.ieee.org/document/7374179?arnumber=7374179>
- Holley, D., & Dobson, C. (2008). Encouraging student engagement in a blended learning environment: The use of contemporary learning spaces. *Learning, Media and Technology*, 33(2), 139–150. Available online from doi.org/10.1080/17439880802097683

- Hughes, G. (2007). Using blended learning to increase learner support and improve retention. *Teaching in Higher Education*, 12(3), 349–363. Available online from doi/abs/10.1080/13562510701278690
- Illinois Online Network. (2008). *Quality Online Course Initiative*. Retrieved from <http://www.ion.uillinois.edu/initiatives/qoci/index.asp>
- IpSOS MORI, & HEFCE. (2016). *About the NSS*. Retrieved from <http://www.thestudentsurvey.com/about.php>
- Khan, B. (n.d.). *e-learning Framework and Models*. Retrieved on 14 April 2016 from <http://asianvu.com/bk/framework/>
- Krause, K. L., & Coates, H. (2008). Students' engagement in first-year university. *Assessment & Evaluation in Higher Education*, 33(5), 493–505. Retrieved from <http://www.tandfonline.com/doi/abs/10.1080/02602930701698892>
- Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education motivation matters. *Educational Researcher*, 41(9), 352–362. Retrieved from <http://edr.sagepub.com/content/41/9/352.short>
- López-Pérez, M. V., Pérez-López, M. C., & Rodríguez-Ariza, L. (2011). Blended learning in higher education: Students' perceptions and their relation to outcomes. *Computers & Education*, 56(3), 818–826. Available online from doi.org/10.1016/j.compedu.2010.10.023
- Ma, C.-M., Chao, C.-M., & Cheng, B.-W. (2013). Integrating Technology Acceptance Model and Task-technology Fit into Blended E-learning System. *Journal of Applied Sciences*, 13(5), 736–742. Available online from doi 10.3923/jas.2013.736.742
- Maryland Online, Inc. (2008). *Quality Matters Rubric Standards 2008–2010 edition with Assigned Point Values*. Retrieved from <https://my.msje.edu/web/ol/ol/RubricStandards2008-2010.pdf>
- Mirriahi, N., Alonzo, D., & Fox, B. (2015). A blended learning framework for curriculum design and professional development. *Research in Learning Technology*, 23. Available online from doi.org/10.3402/rlt.v23.28451
- Office of National Statistics. (2015). *Statistical bulletin: Internet Access – Households and Individuals 2015*. Retrieved from <http://www.ons.gov.uk/ons/rel/rdit2/internet-access--households-and-individuals/2015/stb-ia-2015.html>
- Ozkan, S., & Koseler, R. (2009). Multi-dimensional students' evaluation of e-learning systems in the higher education context: An empirical investigation. *Computers & Education*, 53(4), 1285–1296. Available online from doi.org/10.1016/j.compedu.2009.06.011
- Padilla-Meléndez, A., Del Aguila-Obra, A. R., & Garrido-Moreno, A. (2013). Perceived playfulness, gender differences and technology acceptance model in a blended learning scenario. *Computers & Education*, 63, 306–317. Available online from doi.org/10.1016/j.compedu.2012.12.014
- Pérez, D. P., & Riveros, R. M. (2014). *Unleashing the power of blended learning and flipped classroom for English as Foreign Language learning: Three spheres of challenges and strategies in a Higher Education Institution in Colombia*. Paper presented at the 7th International Conference of Education, Research and Innovation (ICERI) 2014, Seville, Spain. Retrieved from <https://library.iated.org/view/PARRAPEREZ2014UNL>
- Pombo, L., & Moreira, A. (2012). Evaluation Framework for Blended Learning Courses: A Puzzle Piece for the Evaluation Process. *Contemporary Educational Technology*, 3(3), 201–211. Retrieved from https://www.researchgate.net/profile/Lucia-Pombo/publication/234033727_Evaluation_framework_for_blended_learning_courses_a_puzzle_piece_for_the_evaluation_process/links/02e7e52288336c4082000000.pdf
- Pye, G., Holt, D., Salzman, S., Bellucci, E., & Lombardi, L. (2015). Engaging diverse student audiences in contemporary blended learning environments in Australian higher business education: Implications for Design and Practice. *Australasian Journal of Information Systems*, 19. Available online from doi.org/10.3127/ajis.v19i0.1251
- Reed, P. (2014). Staff experience and attitudes towards Technology Enhanced Learning initiatives in one Faculty of Health & Life Sciences. *Research in Learning Technology*, 22. Available online from doi.org/10.3402/rlt.v22.22770
- Saville, N. (2010). Developing a model for investigating the impact of language assessment. *Research Notes*, 42, 2–8. Retrieved from <http://www.cambridgeesol.org/images/23160-research-notes-42.pdf#page=3>
- Shee, D. Y., & Wang, Y.-S. (2008). Multi-criteria evaluation of the web-based e-learning system: A methodology based on learner satisfaction and its applications. *Computers & Education*, 50(3), 894–905. Available online from doi.org/10.1016/j.compedu.2006.09.005
- Smythe, M. (2011). *Blended learning: A transformative process*. Paper presented at the National Tertiary Learning and Teaching Conference 2011, Nelson, New Zealand. Retrieved from <https://akoaootea.ac.nz/download/ng/file/group-3740/smythe---blended-learning-a-transformative-process.pdf>
- So, H.-J., & Brush, T. A. (2008). Student perceptions of collaborative learning, social presence and satisfaction in a blended learning environment: Relationships and critical factors. *Computers & Education*, 51(1), 318–336. Available online from doi.org/10.1016/j.compedu.2007.05.009
- Sparks, S. D. (2015). Blended Learning Research Yields Limited Results. *Education Week*, 34(27), 12–14. Retrieved from <http://www.edweek.org/ew/articles/2015/04/15/blended-learning-research-yields-limited-results.html>
- Stockwell, B. R., Stockwell, M. S., Cennamo, M., & Jiang, E. (2015). Blended Learning Improves Science Education. *Cell*, 162(5), 933–936. Available online from doi.org/10.1016/j.cell.2015.08.009
- Sun, P.-C., Tsai, R. J., Finger, G., Chen, Y.-Y., & Yeh, D. (2008). What drives a successful e-Learning? An empirical investigation of the critical factors influencing learner satisfaction. *Computers & Education*, 50(4), 1183–1202. Available online from doi.org/10.1016/j.compedu.2006.11.007
- Taylor, F. A., Nelson, E., Delfino, K., & Han, H. (2015). A Blended Approach to Learning in an Obstetrics and Gynecology Residency Program: Proof of Concept. *Journal of Medical Education and Curricular Development*, 2015(2), 53–62. Available online from doi: 10.4137/JMECD.S32063
- The Sloan Consortium. (2011). *A Quality Scorecard for the Administration of Online Education Programs*. Retrieved on 14 April 2016 from <http://onlinelearningconsortium.org/consult/quality-scorecard/>
- Tobin, K. (1998). Qualitative Perceptions of Learning Environments on the World Wide Web. *Learning Environments Research*, 1(2), 139–162. Retrieved from <http://link.springer.com/article/10.1023/A:1009953715583>
- Trowler, V. (2010). Student engagement literature review. *The Higher Education Academy*, 11, 1–15. Retrieved from <http://www.lancaster.ac.uk/staff/trowler/StudentEngagementLiteratureReview.pdf>
- Wang, Y.-S. (2003). Assessment of learner satisfaction with asynchronous electronic learning systems. *Information & Management*, 41(1), 75–86. Available online from doi.org/10.1016/S0378-7206(03)00028-4
- Wivell, J., & Day, S. (2015). Blended learning and teaching: Synergy in action. *Advances in Social Work and Welfare Education*, 17(2), 86–99. Retrieved from <https://search.informit.com.au/documentSummary;dn=610028976437941;res=IELHSS>

An analysis of the effect of taking the EPQ on performance in other Level 3 qualifications

Tim Gill Research Division

Introduction

The Extended Project Qualification (EPQ) is a stand-alone qualification taken by students in Year 12 or Year 13, usually alongside General Certificate of Education (GCE) Advanced levels (A levels). It is equivalent in size to half an A level and involves undertaking a substantial project in an area of personal interest, where the outcome can range from writing a dissertation or report to creating a piece of art or organising an event. Its aims are summed up by the following quote from the Department for Education and Skills (DfES) White Paper which proposed the qualification:

This will be a single piece of work, requiring a high degree of planning, preparation, research and autonomous working. The projects will require persistence over time and research skills to explore a subject independently in real depth.
(DfES, 2005, p.63)

One of the perceived benefits of taking the EPQ is that the skills that are learnt by students whilst undertaking their project (e.g., planning, researching, critical thinking, etc.) may be useful for them in their future studies. In the evaluation of the EPQ pilot (Centre for Education and Industry [CEI], 2008), interviews and surveys were used to collect the views of teachers and students in centres offering the qualification, and of representatives from Higher Education (HE). A majority of teachers and HE representatives agreed that the EPQ taught students the skills and competencies that are required for university study and these skills were not assessed in other qualifications. Other research used quantitative analyses of assessment data to show that EPQ grades are good predictors of degree outcomes (Gill & Vidal Rodeiro, 2014) and that students taking the EPQ alongside A levels are more likely to achieve a good degree than those taking A levels only (Gill, 2016a).

It is also of interest to consider whether the skills learnt whilst undertaking the EPQ might be transferable to other qualifications taken at the same time, and might therefore improve performance in those qualifications. In the CEI 2008 study, a majority of teachers surveyed agreed that the EPQ helped their students (at least 'to some extent') with other qualifications taken at the same time. This was due to the new skills the students learnt, and also due to an increase in self-confidence and motivation that came from working independently. On the quantitative side, research by Jones (2015) using data from the AQA exam board found that taking the EPQ alongside A levels increased the odds of achieving a high grade (A* to B) at A level by 29 percent.

The potential for such improvement can also be inferred from research which found positive effects of other qualifications or programmes which explicitly teach or assess skills rather than

knowledge. Black and Gill (2011) found that taking an Advanced Subsidiary (AS) level in Critical Thinking (and achieving at least a grade B) had a positive effect on overall performance at A level that was worth one quarter of a grade on average. Jones, Gaskell, Prendergast and Bavage (2016) found that teaching a pre-university skills course to Year 12 students in order to prepare them better for university study had the (unintended) effect of improving performance in A levels. Finally, a report by Stock Jones, Annable, Billingham and MacDonald (2016) investigated the impact of a programme designed to encourage Science, Technology, Engineering and Mathematics (STEM) participation. The British Science Association's Silver CREST Award gets General Certificate of Secondary Education (GCSE) level students to undertake their own STEM-related projects. Stock Jones et al. (2016) found that students undertaking a CREST project achieved better Science GCSE results by half a grade compared to a control group of statistically matched students who did not do a project.

The main aim of the research reported in this article was to investigate whether students taking the EPQ performed better, on average, in other qualifications compared with their counterparts who did not take the EPQ. This is similar to the analysis undertaken by Jones (2015), but extends it to include data from all exam boards in England and looks at the overall effect by student, rather than by A level entry. It also includes an investigation of the effect of the EPQ at centre level, alongside the student level analysis.

Student level analysis

Data and methods

The data used in the analysis was taken from the National Pupil Database (NPD). This database is managed by the Department for Education (DfE), and consists of all examination results for all pupils in schools and colleges in England, as well as pupil and school background characteristics (e.g., gender, ethnicity, deprivation). For this research, the Key Stage 5 (KS5) datasets for two different years were used. These include all results for students who were aged between 16 and 18 at the end of the academic year, and had taken at least one qualification in the current year equal in size to one A level. They include the results of qualifications taken by these students in previous years, such as AS levels (or the EPQ) taken in Year 12 by students currently in Year 13.

Data from the NPD for 2013/14 and for 2014/15 was used and separate analyses were undertaken for the two different academic years. As most students taking the EPQ combined it with A levels and AS levels or with A levels only (and usually this was a minimum of three A levels), it seemed sensible to make comparisons within this group of students only. Therefore, for the student level analysis,

a subset of NPD data was created, consisting of all students taking at least three A levels combined with at least one AS level or EPQ (or both) and no other qualifications. Qualifications that were retaken were counted only once and the best grade kept.

A multilevel regression model was undertaken for each year, with the outcome variable being the mean University and College Admissions Service (UCAS) points score (excluding the EPQ result, where taken). The effect of candidate ability was accounted for by including a measure of prior attainment in the models (Key Stage 4 [KS4] mean points score, centred on its mean). A further variable was included for the total size of the qualifications taken by a student (including the EPQ, where taken). This was measured in terms of A level equivalents (e.g., A level = 1, AS level = 0.5, EPQ = 0.5). This was an attempt to account for two possible, though opposing, effects: first, a motivation effect whereby students choosing to take more qualifications may be more motivated, leading to them performing better on average; secondly, students taking a large number of qualifications may be over-worked, leading to them performing less well on average. To make interpretation of this variable easier, the minimum size (in this cohort of students) of three and a half (equal to three A levels + one AS level or EPQ) was subtracted from each value. This meant that the baseline for the variable was taking the equivalent of three and a half A levels and the parameter estimate represented the change in the outcome variable associated with taking one more A level (or equivalent).

Two background characteristics were also included: gender and school type. Students were also classified by whether or not they took the EPQ and this was included in the models. A statistically significant parameter estimate for this variable would indicate that taking the EPQ was associated with better (or worse) overall performance in KS5 qualifications. Finally some interaction terms between the EPQ variable and other contextual variables (KS4 mean points score, gender, school type and qualification size) were included to explore whether the effect of taking the EPQ was different for different groups of students.

The hierarchical nature of the data meant that it was appropriate to use multilevel regression models. These take account of the fact that data at one level (students) was 'nested' within another level (schools). Outcomes tend to be more similar within schools than between schools and so to ignore this structure would potentially lead to incorrect results. For a more detailed description of multilevel logistic regressions see Goldstein (2011).

The models presented in this analysis took the following general form:

$$Y_{ij} = \beta_0 + \beta_1 IV1_{ij} + \beta_2 IV2_{ij} + \dots + \beta_k IVk_{ij} + u_j + e_{ij}$$

where Y_{ij} is the mean UCAS points score for student i in school j , $IV1$ to IVk were the independent variables (including the contextual variables and whether or not the student took the EPQ), β_0 to β_k were the regression coefficients, u_j was a random variable at school level and e_{ij} was an individual level residual.

Results

Descriptive

Tables 1 to 3 present descriptive data on the students taking EPQ, compared with those not taking the qualification (i.e., taking A levels only, or A levels combined with AS levels). This shows that EPQ

Table 1: Percentage of EPQ and non-EPQ students in different groups

	2013/14		2014/15	
	EPQ	Non-EPQ	EPQ	Non-EPQ
No. of students	23,396	110,203	24,510	115,731
All	17.5	82.5	17.5	82.5
Female	61.3	55.8	63.4	56.5
Male	38.7	44.2	36.6	43.5
Academy	27.5	28.6	30.9	31.2
Comprehensive	19.0	22.7	19.9	20.8
FE/Tertiary	5.3	7.0	3.8	6.6
Independent	11.4	16.1	11.6	15.4
Other	0.8	0.8	1.6	0.8
Grammar	6.9	4.2	7.6	4.0
Sixth Form	29.2	20.6	24.5	21.2

Table 2: Comparison of EPQ and non-EPQ students, 2013/14

	EPQ		Non-EPQ	
	Mean	SD	Mean	SD
KS4 mean points score	50.1	4.6	48.5	4.5
Qualification size	4.3	0.6	3.8	0.5
Mean UCAS points	96.5	22.7	87.7	23.0

Table 3: Comparison of EPQ and non-EPQ students, 2014/15

	EPQ		Non-EPQ	
	Mean	SD	Mean	SD
KS4 mean points score	49.9	4.6	48.3	4.5
Qualification size	4.3	0.5	3.8	0.4
Mean UCAS points	96.4	22.7	87.7	22.8

students were more likely than non-EPQ students to be female (61.3% in 2013/14 and 63.4% in 2014/15) and to attend sixth form colleges or grammar schools, and were less likely to attend comprehensive or independent schools or Further Education (FE) or Tertiary colleges.

In terms of their prior attainment, EPQ students had a higher average KS4 points score (50.1 compared with 48.5 in 2013/14; 49.9 compared with 48.3 in 2014/15). EPQ students also performed better on average in terms of average UCAS points and tended to have taken more qualifications.

Modelling (2013/14)

The results of the modelling using 2013/14 data are presented in Table 4. The model building proceeded as follows: Model 1 included no predictors, just an intercept, to assess the amount of variance in achievement between schools. From the error variance part of the table we can calculate that around 20.5 percent of the variance was accounted for by schools¹. This is a substantial proportion of the variance and suggests that the use of a multilevel model was justified.

1. As calculated by the intraclass correlation coefficient (ICC). ICC = school variance/(school variance + error variance) = 114.060/(114.060 + 440.160) = 0.205.

Table 4: Model parameter estimates for student level analysis, 2013/14
(standard errors in brackets)

<i>Fixed effects</i>		<i>Model 1</i>	<i>Model 2</i>	<i>Model 3</i>	<i>Model 4</i>
Intercept		87.028 (0.234)	87.403 (0.146)	86.681 (0.214)	86.623 (0.219)
KS4 points score			3.182 (0.012)	3.156 (0.012)	3.120 (0.013)
Gender	Male Female		1.107 (0.102)	1.139 (0.101)	1.587 (0.110)
Qualification size			-1.121 (0.123)	-1.050 (0.123)	-1.335 (0.134)
EPQ	No Yes		5.309 (0.151)	5.239 (0.150)	5.360 (0.361)
School type	Academy Comprehensive FE/Tertiary Independent Other Grammar Sixth Form			-0.374 (0.287) -2.875 (0.523) 5.494 (0.330) -2.100 (1.142) 0.156 (0.636) -0.349 (0.486)	-0.352 (0.293) -3.354 (0.531) 5.585 (0.336) -2.603 (1.151) -0.177 (0.653) -0.640 (0.488)
KS4 points score*EPQ					0.184 (0.030)
Gender*EPQ	Male Female				-2.632 (0.256)
School type*EPQ	Academy Comprehensive FE/Tertiary Independent Other Grammar Sixth Form				-0.225 (0.404) 3.072 (0.660) -0.790 (0.480) 4.259 (1.639) 1.444 (0.637) 1.600 (0.374)
Qualification size*EPQ					1.127 (0.257)
<i>Error variance</i>					
Level 1		440.160 (1.720)	277.980 (1.099)	278.060 (1.099)	227.620 (1.097)
Level 2 – intercept		114.060 (3.899)	27.264 (1.153)	20.970 (0.943)	20.743 (0.934)
<i>Model fit</i>					
AIC		1197428	1109291	1108896	1108689

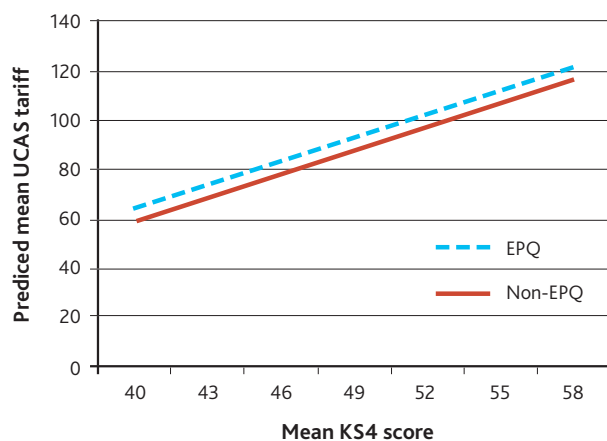
AIC = Akaike Information Criterion

Model 2 then includes the Level 1 predictors (prior attainment, gender, qualification size and whether EPQ was taken). Model 3 adds in the Level 2 predictor (school type) and finally Model 4 adds in interaction terms between the EPQ indicator and each of the other predictor variables. In these, and all subsequent models, statistically significant effects are signified by bold type.

Looking at Model 3 first of all, we can see that there is evidence that taking the EPQ was beneficial to overall performance at KS5. Although the difference of around five points is equivalent to only about a quarter of a grade on average (i.e., one grade in one qualification if taking four A levels), this could still be an important difference in practice. For example, it could mean the difference between meeting

and not meeting a university offer. The other variables in this model were all significant, with females being more likely to do well than males, and students taking more qualifications less likely to do well. Compared to academy schools, students at FE/Tertiary colleges were less likely to do well, whilst those at independent schools were more likely to do well.

To illustrate the magnitude of the EPQ effect, Figure 1 uses the results of Model 3 to compare (at different levels of prior attainment) the predicted UCAS tariff for students taking the EPQ, with the predicted UCAS tariff for those not taking the EPQ. These predictions were for a male student at an academy school, taking qualifications equal to four A levels – either three A levels and two AS levels, or three A levels, one AS level and the EPQ.



Prior attainment	Predicted UCAS tariff (Non-EPQ)	Predicted UCAS tariff (EPQ)
40	59.3	64.5
43	68.7	74.0
46	78.2	83.4
49	87.7	92.9
52	97.1	102.4
55	106.6	111.8
58	116.1	121.3

Figure 1: Predicted UCAS tariff by prior attainment level, EPQ and non-EPQ students, 2013/14, (Model 3)

Table 5: Model parameter estimates for student level analysis, 2014/15

(standard errors in brackets)

Fixed effects		Model 1	Model 2	Model 3	Model 4
Intercept		86.720 (0.231)	87.421 (0.141)	87.221 (0.201)	87.267 (0.205)
KS4 points score			3.143 (0.012)	3.121 (0.012)	3.098 (0.013)
Gender	Male				
	Female		0.796 (0.099)	0.832 (0.099)	1.137 (0.107)
Qualification size			-2.365 (0.125)	-2.322 (0.125)	-2.734 (0.137)
EPQ	No				
	Yes		5.741 (0.148)	5.746 (0.148)	5.072 (0.345)
School type	Academy			-1.012 (0.282)	-1.112 (0.289)
	Comprehensive			-3.816 (0.537)	-4.203 (0.543)
	FE/Tertiary			4.220 (0.323)	4.207 (0.323)
	Independent			-5.603 (1.149)	-5.844 (1.162)
	Other			-0.423 (0.642)	-0.547 (0.661)
	Sixth Form			-0.220 (0.485)	-0.434 (0.487)
KS4 points score*EPQ					0.108 (0.029)
Gender*EPQ	Male				
	Female				-1.845 (0.252)
School type*EPQ	Academy				0.567 (0.404)
	Comprehensive				2.839 (0.660)
	FE/Tertiary				0.007 (0.480)
	Independent				1.868 (1.639)
	Other				0.402 (0.637)
	Sixth Form				1.114 (0.374)
Qualification size*EPQ					1.781 (0.269)
Error variance					
Level 1		433.750(1.653)	277.470 (1.070)	277.500 (1.070)	227.250 (1.069)
Level 2 – intercept		114.570 (3.831)	25.971 (1.100)	21.141 (0.934)	20.925 (0.927)
Model fit					
AIC		1255345	1164179	1163857	1163733
BIC		1255363	1164220	1163933	1163862

AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion

Thus, a male student with a prior attainment of 52 points (equivalent to all grade As) and not taking the EPQ was predicted a mean UCAS tariff of just over 97 points (equivalent to A level grades of BBB and AS levels grades of B and C). If he did take the EPQ the prediction is 102.4 points (equivalent to grades ABB in the A levels and a grade C in the AS level).

Returning to the analysis presented in Table 4, we can see that if we include the interaction terms (Model 4), the effect of the EPQ was again around five UCAS points. However, because of the interaction effects, this EPQ effect only refers to students in the baseline category for all other variables (male students, taking the equivalent of three and a half A levels, attending an academy school, and with a KS4 points score equal to the mean). The interactions between EPQ and KS4 points score, gender, school type and qualification size mean that the effect of the EPQ was found to be different for different levels of each variable. Thus, as KS4 points score increased, the effect of the EPQ became significantly larger, but for female students it was significantly smaller (compared to males). The effect of taking the EPQ was also significantly larger for students in FE/Tertiary colleges, 'Other' schools, grammar schools and sixth form colleges, and for those taking more qualifications.

Modelling (2014/15)

Table 5 presents the results using the 2014/15 data.

The results were very similar to the models using the 2013/14 data. This time schools accounted for 20.9 percent of the variation in the outcomes. The value of the EPQ parameter in Model 3 suggests that taking the EPQ was beneficial to overall performance at KS5, by around five points (equivalent to one grade in one subject if taking four A levels). Other variables were all significant, including females being more likely to do well than males, and students taking more qualifications less likely to do well. Compared to academy schools, students at comprehensive schools, FE/Tertiary colleges or 'Other' schools were less likely to do well, whilst those at independent schools were more likely to do well.

If we include the interaction terms (Model 4 in Table 5) we can see that the effect of the EPQ was again around five UCAS points. However, this effect was only for students in the base category for all variables. The interactions show that, as KS4 points score increased, the effect of the EPQ became larger, but for female students it was smaller (compared to males). The effect of taking the EPQ was also larger for students in FE/Tertiary colleges and sixth form colleges, and for those taking more qualifications.

The results of the modelling were very similar, whether using the 2013/14 or the 2014/15 data. They show that taking the EPQ did have a statistically significant and positive effect on student performance in terms of the UCAS points tariff. However, the effect was quite small, equivalent to around one grade in one A level if taking four A levels.

As a further check on the robustness of these results, two further models were run (using the 2014/15 data only) which only included students with the same volume of qualifications (so that the students being compared were more alike). The data for the first of these models was restricted to those taking qualifications equivalent to four A levels (three A levels and two AS levels, or three A levels, one AS level and the EPQ) and for the second model equivalent to three and a half A levels (three A levels and one AS level, or three A levels and the EPQ). The results of the models are presented in Appendix 1. They show mostly very similar results, with a small but significant EPQ effect.

Centre level analysis

Data and methods

The second part of this research investigated the effect at centre level of taking the EPQ. More specifically, looking at whether increasing the proportion of students taking the EPQ in a centre was associated with better overall performance (in all qualifications). To do this, data from the NPD in two different academic years (2009/10 and 2011/12) was used. This data was chosen because the increase in EPQ entries was particularly large between these two years, up from around 18,700 in 2009/10 to over 33,000 in 2011/12 (Gill, 2016b). A gap of two years was thought to be suitable because inspection of the data found that for many centres the uptake of the EPQ was quite low in the first year of offering the qualification and tended to be much higher in the second year. Furthermore, two years is a short enough period that there should not be too many changes within centres in terms of other factors that might affect attainment.

A difference-in-differences design was used to assess the impact of increasing EPQ uptake. This technique is appropriate for assessing the effect of a reform or the introduction of a new programme or policy (see, for example, Abramovsky, Battistin, Fitzsimons, Goodman, & Simpson, 2011; Belot & Vandenberghe, 2014). The outcome variable in such a model is the difference between some outcome measure before and after the reform or programme is introduced. Comparisons can then be made, in terms of this difference, between those exposed to the new reform/programme and those not exposed.

For this research the 'reform' was the introduction of the EPQ in some centres. The outcome variable was the difference in centre mean UCAS tariff between before (2009/10) and after (2011/12) introducing the EPQ. This variable was calculated by adding up the UCAS tariff for each grade achieved in Level 3 qualifications in the centre and dividing by the total size of qualifications taken. The EPQ and any qualifications worth less than half an A level were excluded from this calculation.

The centres included in the models were only those with zero, or very low (less than 5 percent) EPQ uptake in 2010, so that the effect of the introduction of the EPQ into centres which had not previously offered the qualification could be investigated. The inclusion of centres with very low uptake in 2009/10, as well as those with zero uptake, was necessary to boost the number of centres available for the modelling. Only centres whose mean UCAS tariff (in both 2009/10 and 2011/12) was based on at least 20 students were included. This meant that the final dataset for the models included 1,730 centres.

A standard difference-in-differences model would include a binary indicator of whether or not the centre had introduced the EPQ. However, further inspection of the data found that most of the centres introducing the EPQ only had a very low percentage of their students taking the qualification in 2011/12, which is unlikely to have a big effect on outcome measures. This is shown in Figure 2, which presents the distribution of EPQ uptake amongst centres.

To take account of this, the variable indicating introduction of EPQ was split into four separate categories depending on what proportion of students took the EPQ in 2011/12. These categories indicated *zero uptake* (actually less than 5%), *low uptake* (5–10%), *moderate uptake* (10–30%) or *high uptake* (>30%) of the EPQ.

Several centre level contextual variables were included in the models. These were a measure of the average prior attainment of students at the school (KS4 mean points score), the mean size of the qualifications taken

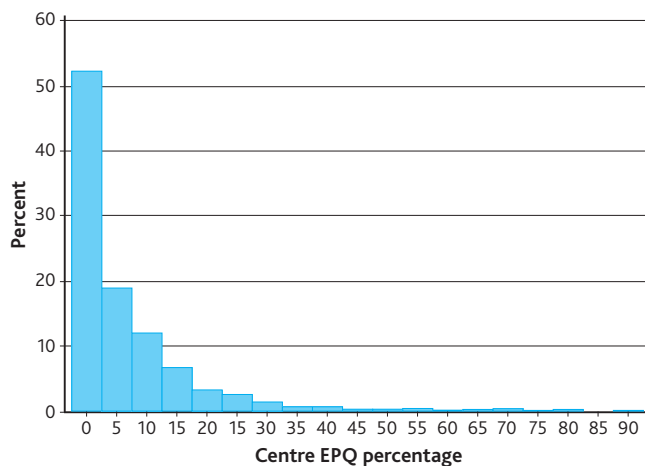


Figure 2: Distribution of centre EPQ uptake in 2011/12

by students, the percentage of white students, the percentage of students eligible for free school meals (FSM) and the school type. For the ethnicity and FSM variables, there was a relatively large amount of missing data, so the percentage of missing data was calculated for each centre and also included in the models.

Finally, to account for any changes within centres between 2009/10 and 2011/12, a difference variable was calculated for each of the contextual variables, equal to the 2011/12 value minus the 2009/10 value (e.g., FSM percentage in 2011/12 – FSM percentage in 2009/10).

Thus, the final models took the following form:

$$\Delta Y_j = (Y_{j2012} - Y_{j2010}) = \beta_0 + \beta_1 IV1_j + \beta_2 IV2_j + \dots + \beta_k IVk_j + u_j$$

where ΔY_j is the change in the mean UCAS tariff for school j between 2010 and 2012, $IV1$ to IVk were the independent variables (including the EPQ category, contextual variables and the variables accounting for differences in contextual factors over time), β_0 to β_k were the regression coefficients and u_j was the residual.

Results

Descriptive

In total there were 1,730 centres included in the model. The distribution of centres by EPQ category was as follows:

Table 6: Distribution of centres by EPQ uptake category

EPQ category	Uptake levels	No. of centres
No uptake	<5%	1,096
Low uptake	5–10%	267
Moderate uptake	10–30%	306
High uptake	>30%	61

Table 7 presents descriptive data on the outcome variable for the models.

Thus, centres in 2012 performed slightly better on average on the measure of attainment. The biggest difference in a centre was about 40 points, equivalent to two A level grades.

Table 7: Descriptive data for difference in centre mean UCAS tariff between 2009/10 and 2011/12

Mean	SD	Min	Max
0.27	7.06	-38.87	44.04

Modelling

Linear regression models were used for this analysis. The only predictor in the first model was the EPQ category. The second model added in the contextual variables and the 'difference' variables. Only variables with statistically significant effects were included in these final models. The results of the models are presented in Table 8.

Model 1 included only the EPQ category as a predictor variable, and showed that centres that introduced the EPQ with at least 30 percent of students had a significantly larger improvement in their mean UCAS tariff between 2009/10 and 2011/12 than centres with no uptake. However, there was no such effect if the EPQ uptake was low or moderate in 2011/12.

The results after including the covariates that were statistically significant (Model 2) show that having low uptake did not make a significant difference, but having moderate or high uptake was associated with a larger increase in the mean UCAS tariff for a centre. The difference was small, just one UCAS point for moderate uptake and two UCAS points for high uptake. Two UCAS points is equivalent to 1/10th of an A level grade. In other words, the model predicts that introducing the EPQ into a centre (with 30 percent or more students taking the qualification) would increase a centre's attainment by one grade for every ten A levels taken, compared with centres not introducing the EPQ.

Although not the main focus of this research, it is interesting to note the effects of the contextual and 'difference' variables included in the model. The only contextual variable that was statistically significant in Model 2 was the percentage of FSM students in the centre². This was negative, indicating that having a higher proportion of FSM students was associated with lower attainment in 2011/12 compared with 2009/10. There were three other statistically significant variables, which indicated the effect of changes within centres between the two years (KS4 mean points score, mean qualification size, and the percentage of female students in the centre). All of these were positive. The positive effect of the change in the mean KS4 points score makes sense intuitively, in that

Table 8: Model parameter estimates for centre level analysis

(standard errors in brackets)

Fixed effects		Model 1	Model 2
Intercept		-0.101 (0.213)	-0.774 (0.275)
EPQ category	None		
	Low	0.268 (0.481)	0.284 (0.438)
	Moderate	0.344 (0.456)	0.920 (0.416)
	High	2.003 (0.928)	1.949 (0.844)
FSM %			-0.069 (0.022)
FSM % missing			-0.004 (0.005)
Mean KS4 points score difference			2.001 (0.117)
Mean qualification size difference			2.233 (0.443)
% of female students difference			0.037 (0.018)
Model fit			
Adjusted R Square		0.003	0.179

2. The percentage of missing FSM was also included in the models despite not being statistically significant because this varied considerably between centres and so could potentially impact on the FSM percentage variable.

if a centre attracts more able students, it is likely to improve its overall performance. Increasing mean qualification size, or the percentage of female students were both associated with larger improvements in attainment in 2011/12 compared with 2009/10.

It is interesting to note the low value for the adjusted R square in Model 2 (0.179), meaning that only around 18 percent of the variability in the outcome variable was explained by the predictor variables. In other words, most of the variability was explained by other factors, which were not included in the model.

Discussion

There is evidence from prior research about the benefits of taking the EPQ, in terms of teaching students the thinking skills and independent learning that may help them prepare for university study (see for example, CEI, 2008; Gill & Vidal Rodeiro, 2014; Gill, 2016a). Undertaking a project based qualification is also associated with improved performance in concurrent GCSE/A level studies in particular circumstances (CEI, 2008; Stock Jones et al, 2016; Jones, 2015). The purpose of the research presented here was to investigate whether taking the EPQ could be advantageous for students in qualifications taken at the same time. This article extends beyond prior work by including data from all (rather than one) exam boards, and conducting student and centre level analysis. It is worth noting that the type of work that the EPQ prepares students for (e.g., research, independent thinking, etc.) is present to a lesser degree in A levels than it is at undergraduate level. However, this is not to say that some of these skills are not useful at A level as well.

The main conclusion from this research is that there was some evidence that taking the EPQ may be beneficial in terms of performance on other qualifications, both at the student level and at the centre level. However, in both cases the effect was relatively small. At the student level, taking the EPQ was associated with an improvement in mean UCAS tariff of around five to six points (in both 2013/14 and 2014/15). This is equivalent to an improvement of one grade in one A level for a student taking four A levels. At the centre level, increasing the EPQ uptake from less than 5% of sixth formers to over 30% between 2010 and 2012 was associated with an increase in the overall performance in a centre. This increase amounted to one tenth of an A level grade (in other words, one grade improvement in every tenth A level taken at the centre).

Although neither of these effects could be considered large, they are still important in practice, when considering that they could be the difference between meeting and failing to meet a university offer.

At the student level there were also some interesting (although small) interaction effects between taking the EPQ and other contextual variables. First, the effect of taking the EPQ was higher for those with higher prior attainment, suggesting that the EPQ may benefit the brightest students most. The effect of the EPQ was also greater for male students than for female students, which contrasts with the overall effect of gender on performance according to the models, which favoured females. Indeed the gender interaction effect was larger than the main gender effect, which means that although the non-EPQ females were predicted a higher mean UCAS than the non-EPQ males, the EPQ females were predicted a lower mean UCAS than the EPQ males. Finally, students attending FE/Tertiary colleges had the biggest improvement in performance from taking the EPQ, compared with not taking it.

In terms of the overall effect at the student level we should be somewhat cautious in the interpretation, because we cannot say for certain that there is a causal relationship. For instance, it may be that students taking the EPQ are more motivated to do well academically than those not doing so and it is this, rather than taking the EPQ per se, that enables them to do better in their A levels.

In the centre level model, the outcome variable was the difference in performance over a period of two years. However, it may be that any positive impact of introducing the EPQ into a centre is less in the first few years, as teachers become familiar with teaching the qualification. This hypothesis is borne out by the evaluation of the EPQ pilot (CEI, 2008) which found that teachers reported that it took time for them to get used to the requirements of the new qualification. Therefore the effect found in the results presented in this article may be an underestimate of the longer term effect. One way of assessing whether the effect increases as centres become more experienced would be to re-run the student level models and include a variable indicating, for each student, how long their centre had been teaching the EPQ.

One factor that has not been explored in this research is the effect of the grade received in the EPQ by students. Black and Gill (2011) found that the overall positive effect of taking AS level Critical Thinking was greater for those who achieved a higher grade in the qualification. It would be interesting to see whether the students who achieved best in their EPQ were those who also did well at A level (after accounting for ability). A further centre level analysis could be undertaken to investigate this, by including the centre level EPQ performance in the models. This might indicate that centres where students do particularly well at the EPQ might be able to improve their overall performance more than centres that do less well in the EPQ (i.e., the EPQ is beneficial, but only if it is taught well).

Another area that might be interesting to explore is whether the EPQ is more beneficial for some A level subjects than for others. Jones (2015) found the positive effect of taking the EPQ on A level performance was present (and very similar in terms of size) for all subject groups apart from Mathematics and Languages, for which there was no significant effect. Research by Gill (2016b) found that correlations between the EPQ grade and A level grades differed depending on the A level subject, with the best correlations (amongst the top 10 most common A levels taken by EPQ students) for English Literature (0.47) and History (0.47), and the worst for Mathematics (0.37) and Sociology (0.38). This suggests that the skills learned in the EPQ may be more applicable to some subjects than to others.

References

- Abramovsky, L., Battistin, E., Fitzsimons, E., Goodman, A., & Simpson, H. (2011). Providing Employers with Incentives to Train Low-Skilled Workers: Evidence from the UK Employer Training Pilots. *Journal of Labour Economics*, 29(1), 153–193. Available online from doi: 10.1086/656372.
- Belot, M., & Vandenbergh, V. (2014). Evaluating the 'threat' effects of grade repetition: exploiting the 2001 reform by the French-Speaking Community of Belgium. *Education Economics*, 22(1), 73–89. Available online from doi: 10.1080/09645292.2011.607266
- Black, B., & Gill, T. (2011). Does doing Critical Thinking AS level confer any advantage for candidates in their performance on other A levels? *Research Matters: A Cambridge Assessment publication*, 11, 22–25. Available online at: <http://www.cambridgeassessment.org.uk/Images/109987-research-matters-11-january-2011.pdf>

- CEI (2008). *Evaluation of the Extended Project Pilots: Final Report for the Qualifications and Curriculum Authority (QCA)*. Coventry, UK. Centre for Education and Industry, University of Warwick.
- DfES (2005). *14–19 Education and Skills* (White Paper). London. The Stationery Office. Retrieved from <http://www.educationengland.org.uk/documents/pdfs/2005-white-paper-14-19-education-and-skills.pdf>
- Gill, T. (2016a). *A statistical comparison of level 3 qualifications in England as preparation for university*. Manuscript submitted for publication.
- Gill, T. (2016b). *Uptake and results in the Extended Project Qualification 2008–2015*. Statistics Report Series No.101. Cambridge, UK: Cambridge Assessment. Available online at: <http://www.cambridgeassessment.org.uk/Images/306859-uptake-and-results-in-the-extended-project-qualification-2008-2015.pdf>
- Gill, T., & Vidal Rodeiro, C. L. (2014). *Predictive validity of level 3 qualifications: Extended Project, Cambridge Pre-U, International Baccalaureate, BTEC Diploma*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. Available online at: <http://www.cambridgeassessment.org.uk/Images/178062-predictive-validity-of-level-3-qualifications.pdf>
- Goldstein, H. (2011). *Multilevel Statistical Models (4th edition)*. Chichester: John Wiley & Sons.
- Jones, B. (2015). *Does the Extended Project Qualification enhance students' GCE A-level performance?* Manchester: AQA Centre for Education Research and Practice. Retrieved from https://cerp.aqa.org.uk/sites/default/files/pdf_upload/CERP_TR_BEJ_21052015.pdf
- Jones, H.L., Gaskell, E.H., Prendergast, J.R., & Bavage, A.D. (2016). Unexpected benefits of pre-university skills training for A-level students. *Educational Studies* 43(1), 67–70. Available online from doi: 10.1080/03055698.2016.1245601
- Stock Jones, R., Annable, T., Billingham, Z., & MacDonald, C. (2016). *Quantifying CREST: What impact does the Silver CREST Award have on science scores and STEM subject selection?* London, UK. British Science Association. Retrieved from <http://www.britishsocietyassociation.org/crestsilver-report>

Appendix

This analysis checks the results of the student level modelling by running the same models on a subset of students (for 2014/15 only) with the same volume of qualifications; students taking three A levels and two AS levels, or three A levels, one AS level and the EPQ, and then three A levels and one AS level or three A levels and the EPQ.

Table A1 presents the numbers of students in each group. Table A2 presents the results of the model with UCAS tariff as the outcome variable.

The results are very similar to the model with the full data, with the EPQ effect being slightly higher for the two AS levels model than for the one AS level model. Interestingly, for both models the EPQ effect decreased as KS4 increased, which is the opposite of the effect in the original model.

Table A1: Number of students taking each combination of qualifications

Combination (A level + AS level + the EPQ)	Number of students
3 + 2 + 0	18,582
3 + 1 + 1	13,609
3 + 1 + 0	78,451
3 + 0 + 1	1,900

Table A2: Model parameter estimates for student level analysis on subsets of students, 2014/15

(standard errors in brackets)

Fixed effects		3+2+0 v 3+1+1	3+1+0 v 3+0+1
Intercept		85.912 (0.353)	85.481 (0.224)
KS4 points score		3.201 (0.030)	2.945 (0.016)
Gender	Male		
	Female	0.938 (0.260)	1.392 (0.131)
EPQ	No		
	Yes	6.610 (0.487)	3.028 (0.474)
School type	Academy		
	Comp	-2.438 (0.504)	-0.691 (0.313)
	FE/Tertiary	-3.796 (0.871)	-3.110 (0.574)
	Independent	3.566 (0.633)	4.826 (0.352)
	Other	-3.648 (1.974)	-3.938 (1.282)
	Grammar	-1.117 (0.970)	0.322 (0.799)
Sixth Form		-0.317 (0.664)	-0.159 (0.506)
KS4 points score*EPQ		-0.112 (0.047)	-0.686 (0.093)
Gender*EPQ	Male		
	Female	-0.987 (0.402)	n.s.
School type*EPQ	Academy		
	Comp	1.697 (0.665)	n.s.
	FE/Tertiary	3.026 (1.073)	n.s.
	Independent	1.121 (0.815)	n.s.
	Other	-1.852 (2.434)	n.s.
	Grammar	0.725 (1.211)	n.s.
Sixth Form		0.555 (0.641)	n.s.

A review of instruments for assessing complex vocational competence

Jackie Gcreatorex, Martin Johnson and Victoria Coleman Research Division

Introduction

Complex competences integrate a variety of skills. For example, models of professional proficiency or intelligent practice often incorporate the ability of a person to construct a holistic view of a problem or situation (Dreyfus & Dreyfus, 1980; Eraut, 1994). There is evidence that observational methods can be used to capture the integration of skills, knowledge and attitudes that pertain to higher level work (Eraut & Steadman, 1998). The aim of this research was to explore the measurement qualities of prevalent approaches to observation (checklists and Global Rating Scales [GRSs]) in the context of assessing complex competence.

According to Lester (2000), the assessment of complex competence is possible if the performances assessed are approached holistically rather than in an instrumental or piece-by-piece fashion. At the same time this presents a challenge to assessment. Watson (1994) argues that for observation-based assessment to be reliable, fair, generally practicable and cost-effective it needs to include adequate quality control to ensure consistency across assessors, and involve sensible decisions about the range and number of observations of performance that are required to make a reliable judgement about competence.

Since the holistic assessment of complex competence is a challenge, it is useful to look more closely at cases where assessment models are used to capture complex competence and are considered to be trustworthy. Training in the medical field is a safety-critical professional context involving the assessment of important competences. Moreover, these assessment processes are highly respected since they result in the certification of practice in a very high-stakes professional domain. This article looks more closely at aspects of the assessment processes used in this context to explore how observation-based assessment is used to assess complex competences without concern that the assessments compromise validity. The discussion of assessment of complex competence is foregrounded with a review of human judgement research.

One assessment approach is the checklist approach. This approach involves the development of a checklist of features that are used as the basis for observations of performance. Checklists require raters to indicate the performance or omission of directly observable actions with a separate checklist required for each task (Ilgen, Ma, Hatala, & Cook, 2015). The items are scored for presence or absence.

A concern with this approach is that it leads to an atomistic construction of competence, which narrows the scope of initiative and field of responsibility of professional practitioners, and fails to encompass matters such as maturity, critical thinking, group work, and complex skills (Winter, 1995). In addition, since assessment can be a significant influence on learning, such an assessment approach could also lead to the construction of learning situations where the notion of simple competence dominates.

A second approach is the use of a GRS. These scales require raters to judge participants' overall performance or to provide impressions of performance on subtasks or traits (Ilgen et al., 2015; Norcini, 2005). There is no rule about how many points should be in the scale. The GRS is applied to several traits, such as physical examination and history taking, as well as in several situations such as in Accident and Emergency, and in General Practice. One distinguishing characteristic of a GRS is it is used for multiple situations and traits. It must be noted however that the term 'global rating' is not always used consistently, with greater clarification needed across research in how it is defined and distinguished from other scoring instruments (Boursicot et al., 2011).

In the absence of examples of tasks with mark schemes in the form of checklists and a GRS, we developed the following fictional assessment tasks and extracts from mark schemes for illustrative purposes only. The first task involved the candidate leading a meeting about the progress of a project. The second task involved the candidate giving a 15-minute presentation about the completed project to a group of 20 peers who have not been involved in the project. Both performances were to be observed and rated. The fictional mark schemes provided are:

- GRS for use with Task 1 and 2 (Table 1 on page 36)
- Checklist for use with Task 1 (Table 2 on page 36)
- Checklist for use with Task 2 (Table 3 on page 36).

Human judgement

There is a literature about the strengths and weaknesses of using human judgement. Here we offer a short exploration of the pros and cons of the use of GRSs and checklists to inform human judgement.

GRS and human judgement

Research has indicated that GRSs can give an accurate overview of students' abilities. For instance, the surgery skills of medical students were assessed through ratings on a GRS on ten specific traits (Pulito, Donnelly, & Plymale, 2007). Students were also assigned a grade summarising their performance which was based on the examiner's perception of the student's overall performance, considering any additional factors, and weighting their performance of the ten traits as they felt was appropriate. It was found that the rating on any of the ten specific traits was 75–80 percent accurate in predicting a student's overall grade. Thus this shows that scores on a GRS were able to accurately reflect judgements of students' overall performance. However, it also indicates that examiners tend to make single overall judgements on a student's performance rather than considering each trait separately, suggesting that using multi-item GRSs is unnecessary. That said, there was evidence of some variation between traits with the non-cognitive aspects rated higher overall compared to cognitive aspects. The limitation of this and similar studies is that there is no

Table 1: Fictional GRS for use with Assessment Tasks 1 and 2

Rate the candidate's performance as *Unacceptable*, *Improvement needed*, *Adept*, *Very Good* or *Outstanding* for each of the following traits:

	Level of performance				
	<i>Unacceptable</i>	<i>Improvement needed</i>	<i>Adept</i>	<i>Very Good</i>	<i>Outstanding</i>
	1	2	3	4	5
People management Negotiating allocation of tasks and resources to appropriate staff. Rewarding achievement, giving credit where due and challenging underperformance. Maintaining good relationships.					
Time management Ensuring activities meet deadlines and fit allocated time windows.					
Branding Communications are brand appropriate.					
Written communication Text is clear, succinct and engaging. Sentences and paragraphs are well constructed and build up to an overall conclusion. Text is augmented by varied and imaginative images, graphics and other media which reinforce the message. Images, graphics and other media are accessible and appropriately labelled.					
Expertise Using facts and credible evidence to inform analysis and evaluation which are used to draw conclusions. The content is original. No content is sexist, racist, ageist, homophobic or inflammatory in nature.					
Overall performance					

Table 2: Fictional checklist for use with Assessment Task 1

Tick items which were achieved. All items must be present to gain a pass.

Trait	Meeting
Time management	<input type="checkbox"/> Started and ended on time <input type="checkbox"/> Each section started and ended on time <input type="checkbox"/> The purpose(s) of the meeting was/were clear <input type="checkbox"/> Appropriate timespans were given to each agenda item <input type="checkbox"/> Meeting papers, agenda, minutes of previous meeting were received well before the meeting <input type="checkbox"/> All agenda items were covered in the meeting <input type="checkbox"/> Project activities/stages met deadlines <input type="checkbox"/> Work progress was checked against milestones <input type="checkbox"/> Necessary changes to timelines were made
Branding	<input type="checkbox"/> Organisational template was used <input type="checkbox"/> All images/text and so on met brand guidelines <input type="checkbox"/> Copyright permissions gained as necessary <input type="checkbox"/> Copyright notice added as required
Written communication	<input type="checkbox"/> Tables/figures images/graphics were accessible and augmented the message <input type="checkbox"/> The text was grammatically accurate <input type="checkbox"/> The text was correctly spelt <input type="checkbox"/> Paragraphs had an introduction to the topic, gave evidence about the topic and had a concluding sentence, as appropriate
People management	<input type="checkbox"/> Active listening was exercised <input type="checkbox"/> Questions were answered <input type="checkbox"/> Appropriate responses were given to questions and comments <input type="checkbox"/> Credit was attributed where due <input type="checkbox"/> All meeting attendees had the opportunity to contribute as relevant <input type="checkbox"/> Discussion focused on the issues to hand <input type="checkbox"/> All relevant perspectives were considered before agreeing a way forward
Expertise	<input type="checkbox"/> Expert knowledge was demonstrated <input type="checkbox"/> Conclusions were drawn via analysis of facts or evaluation of evidence <input type="checkbox"/> Content was devoid of sexist, racist, ageist, homophobic or inflammatory content

Table 3: Fictional checklist for use with Assessment Task 2

Tick items which were achieved. All items must be present to gain a pass.

Trait	Presentation
Time management	<input type="checkbox"/> Started and ended on time <input type="checkbox"/> Each section started and ended on time <input type="checkbox"/> The purpose(s) of the presentation was/were clear <input type="checkbox"/> Actions and deadlines/milestones were agreed and recorded
Branding	<input type="checkbox"/> Organisational template was used <input type="checkbox"/> All images/text and so on met brand guidelines <input type="checkbox"/> Copyright permissions gained as necessary <input type="checkbox"/> Copyright notice added as required <input type="checkbox"/> Organisational authorisation gained as needed
Written communication	<input type="checkbox"/> Tables/figures/images/graphics were accessible and augmented the message <input type="checkbox"/> The text was grammatically accurate <input type="checkbox"/> The text was correctly spelt <input type="checkbox"/> Paragraph had an introduction to the topic, gave evidence about the topic and had a concluding sentence, as appropriate
People management	<input type="checkbox"/> Active listening was exercised <input type="checkbox"/> Questions were answered <input type="checkbox"/> Appropriate responses were given to questions and comments <input type="checkbox"/> Credit was attributed where due <input type="checkbox"/> Questions and comments were requested <input type="checkbox"/> Attendees were attentive
Expertise	<input type="checkbox"/> Expert knowledge was demonstrated <input type="checkbox"/> Conclusions were drawn via analysis of facts or evaluation of evidence <input type="checkbox"/> Content was devoid of sexist, racist, ageist, homophobic or inflammatory content

independent measure against which to compare individual trait scores and overall judgements and thereby determine which is more accurate.

Furthermore, a comparison of a single-trait GRS with a multi-trait GRS found that whilst there was significant correlation between the two, a single-trait GRS was not able to reflect the differences found between different traits, such as the finding that ratings tended to be higher on humanistic traits compared to technical ones (Domingues, Amaral, & Zeferino, 2009). Additionally, the ratings on technical traits correlated particularly well with the single-trait GRS scores. This demonstrates that certain traits may have a greater impact on single-trait GRS scores, and that single-trait GRSs are limited as they cannot reflect variation within performance on specific traits (Domingues et al., 2009). This finding may be due to the psychological phenomenon that people can be good at judging individual traits and less good at combining information into an overall judgement (Einhorn, 1972; Laming, 2004).

Overall, it appears that a GRS can be used to generate scores for specific traits which reflect judgements of the student's overall performance. Additionally, the use of a multi-item GRS enables a more in-depth understanding, although examiners do often give fairly uniform responses across these (Pulito et al., 2007).

Checklists and human judgement

Prior research shows that experts can successfully identify the characteristics in a checklist, but they are poor at combining the decisions from each point in the checklist into an overall judgement. For instance, Eining, Jones, and Loebbecke (1997) evaluated the effectiveness of cue processing aids in fraud detection. The aids were:

- A checklist
- A statistical model (using data collected by humans using a checklist)
- An expert system (using data collected by humans using a checklist)
- Unaided judgement (when auditors make an overall judgement using the evidence available).

The most superior fraud assessment was achieved by the statistical model and the expert system; here unaided judgement was inferior but better than using the checklist alone. Later, Boritz and Timoshenko (2014) reviewed related studies and argued that humans can effectively respond to each item on a checklist, but that mechanical combination of the decisions on each checklist item (statistical model/expert system) is superior to human combination of the decisions on each point on the checklist. Additionally, it is noteworthy that all forms of cue processing aids rely on high-quality checklists which contain all the key traits (Boritz & Timoshenko, 2014).

Combining different types of evidence

There is a host of research about how humans integrate evidence from several sources to make a judgement and the quality of those judgements – examples include Kahneman (2011) and Laming (2004). Here we focus on work comparing human and mechanical approaches to integrating evidence.

Highhouse and Kostek (2013) reviewed research on college admissions and employee selection in the US. Generally the studies compared predictions of college success or achievement in a job from:

- Human integration of information into an overall judgement
- Mechanical integration of evidence.

An illustrative example is that in a police assessment centre each assessor scored each candidate's performance on each exercise, and the assessors jointly provided an overall rating for each candidate (Feltham, 1988). A statistical combination of the scores on various exercises was a better predictor of success as a police officer than the consensus overall judgement. In four of the seven studies about college admissions, a mechanical combination of evidence outperformed human integration of evidence (Highhouse & Kostek, 2013). In 6 of the 13 studies about employee selection, a mechanical combination of evidence outperformed human integration of evidence, and 3 gave the reverse result. Together the research shows that mechanical combination of evidence tends to be better than human judgements which integrate a variety of evidence.

Methods for mechanically combining assessment outputs (scores, grades etc.) are many and varied. An example follows by way of illustration. Many (post)graduate degrees assess aspects of complex competence at many intervals (Janssen et al., 2016). In the case of Medicine these can involve scores, grades (and equivalent) as well as the textual comments of the assessors. One way of combining the scores, grades and text is a *Multi-Entity Bayesian network* (Janssen et al., 2016). A *Bayesian network* is a statistical model that uses Bayesian methods to estimate the parameters of the posterior distribution (probability distribution of an unknown quantity treated as a random variable conditional on the data provided). A Multi-Entity Bayesian network goes beyond Bayesian networks to form complex situation-specific Bayesian networks, and as more data is accrued in the database the network and outputs are updated. In other words, the Multi-Entity Bayesian network can account for the assessment context, which other types of Bayesian models cannot. The data fed to the model are scores, grades (equivalents) and sentiment levels derived from a sentiment analysis of assessors' textual comments (Janssen et al., 2016). The Multi-Entity Bayesian network combines the information and estimates the true present level of performance. Output from the model is posterior probability tables for multiple variables, such as level of motivation. These analytics are interpreted by experts to make decisions about degree classifications, learning needs to be addressed and so on.

Why can mechanical combination be better than human combination of evidence?

To explain why mechanical combination can outperform human integration of evidence we return to theory. Kahneman (2011) explains that there are two reasoning systems controlling human judgement. *System 1* is intuitive, unconscious, automatic and fast. System 1 thinking associates new information with established thought patterns and understandings, rather than noting the uniqueness of the current situation. For example, when a doctor encounters a case of measles and uses System 1 thinking he/she recalls cases he/she previously experienced rather than recognising the distinguishing characteristics of this case. System 1 thinking quickly amalgamates new information into a model (script/schema) based on prior experience, and potentially overlooks key new data. *System 2* thinking is deliberate, conscious, laboured and slow. System 2 thinking integrates information using a coherent judgement model, and can be used to make considered and logical decisions. System 1 thinking often obstructs System 2 thinking, which may influence the quality of human judgement. Both systems must be useful otherwise they would have disappeared through evolutionary processes.

Arguably, human judgement involves simplifying heuristics (Gilovich, Griffin, & Kahnemann, 2002). Generally these heuristics are helpful and provide accurate judgements, however, they can lead to unintentional biases (Gilovich & Griffin, 2002). For example, Tversky and Kahneman (1982) found that sometimes people appraise the likelihood of an event by the ease with which incidences can be recalled. This mental short cut is known as the *availability heuristic*. Often the availability heuristic is successful because recurring events are brought to mind more effortlessly than infrequent events. But the availability heuristic can result in biased judgements, for example, biases due to the retrievability of instances. One group might be judged larger than another, even though the two groups are of equal size. The bias occurs because the group of familiar instances is more easily brought to mind and therefore seems larger.

This theory applies to the situation of an assessor combining performance evidence to give an overall performance rating. The assessor's experience might be that students who are good at physical examinations are sound doctors. That is, the assessor has a script that students who are reasonable at physical examinations are able doctors. Therefore, when the assessor is integrating evidence from physical examinations, professionalism and so on, they give greatest weight to the students' performance on the physical examination. In other words, they used System 1 thinking. If the assessor's experience is correct, then the System 1 thinking was successful. If however, the assessor's schema were factually incorrect, then the System 1 judgement is biased. Some mark schemes might circumvent such biases by requiring assessors to judge each trait separately and then judgements are mechanically combined to give an overall score.

In the following sections we consider how these issues extend into assessing complex competence by focusing on a widely used GRS (*mini-CEX*) and an area where checklists are popular (essential skills).

The mini-CEX: A Global Rating Scale

The *Clinical Evaluation Exercise* (CEX) was designed as a practical assessment of trainee doctors' clinical skills (Norcini, Blank, Duffy, & Fortna, 2003). The CEX involved trainees carrying out a two-hour full history and physical examination of an inpatient, being observed and assessed on their clinical skills by a supervising clinician using a GRS. Whilst the CEX enabled the assessment of a trainee's clinical skills with a real patient, it had limited generalizability beyond this specific context, only involved a single assessor, and was not representative of normal doctor-patient interactions (Norcini, 2005).

The mini-CEX is a modification of the original CEX that was developed by the American Board of Internal Medicine and has since been used in a variety of countries including the UK. It is a GRS which assesses the clinical skills of trainee doctors across a number of settings and scenarios (Norcini, 2005). It involves a higher trained physician assessing the trainees' performance of clinical skills on a routinely conducted clinical task. Trainees are assessed on seven domains: history taking; physical exam; professionalism; clinical judgement; communication skills; organisation/efficiency; and overall clinical care (Norcini et al., 2003). This is done on a rating scale, which ranges from six to nine points with the bottom of the scale representing unsatisfactory/below expectations and the top of the scale superior/above expectations. Assessors are often required to complete this form online and to provide feedback to

trainees immediately afterwards, noting particular strengths or weaknesses (Norcini, 2005). The mini-CEX lasts approximately 15–20 minutes and is carried out during normal clinical activities. Six are carried out in each of the first and second year of the UK foundation programme for trainee doctors. They are organised by the trainee doctors themselves, are spaced out throughout the year, and conducted by a variety of different assessors in different scenarios and settings (Norcini, 2005). It is not necessary for all domains to be assessed on each mini-CEX if they are not relevant to particular scenarios.

When all six mini-CEX are completed the data is collated and returned to the trainee. It was designed as a formative learning and development tool, enabling trainees to reflect on their strengths and weaknesses, rather than having a summative function of measuring proficiency levels, and was not intended as a tool to compare trainees (Norcini, 2005; Weston & Smith, 2014; Yates, 2013). That said, it is frequently used in a summative manner (Hawkins, Margolis, Durning, & Norcini, 2010).

Research evidence about the mini-CEX

The mini-CEX shows good feasibility and can form part of normal clinical practice (Pelgrim, 2010; Yates, 2013). There is evidence that assessors view the mini-CEX favourably (Norcini et al., 2003). Whilst one literature review suggested there was evidence of learner engagement in the mini-CEX (Yates, 2013), others found that trainees did not consider it a useful part of their training which may relate to a lack of understanding of its formative purpose (Weston & Smith, 2014).

Evidence that individuals' mini-CEX rating scores in all domains appear to increase over time is supportive of its construct validity (Hawkins et al., 2010; Pelgrim, 2010). However, factors beyond clinical competence may also influence mini-CEX ratings. Assessors make social judgements when assessing trainees and differences in these judgements impact rating scores (Gingerich, van der Vleuten, Eva, & Regehr, 2014). Differences have been found in the ratings given by assessors who were residents (doctors holding certain degrees who are not yet fully licensed) compared to those who were faculty members (Al Ansari, Ali, & Donnon, 2013). Generalizability of the mini-CEX results may be limited by the influence of examiner factors on reliability, with examiner factors accounting for 23–40% of variance compared to trainee ability which accounts for 4–17% of variance (Yates, 2013).

There is also evidence that, when assessing clinical competence in a real-life setting such as this or in more complex situations, assessors may give overinflated rating scores thus limiting validity (Hawkins et al., 2010; Norcini et al., 2003). Evidence of criterion validity has been inferred by comparing the mini-CEX with other assessment of clinical skills, such as oral and written exams or performance evaluations (Al Ansari et al., 2013; Hawkins et al., 2010; Pelgrim, 2010).

Finally, research suggested that mini-CEX scores from ten encounters produces good reliability (Norcini et al., 2003). Inter-rater reliability of the mini-CEX is influenced by the number of points on a scale, with greater inter-rater reliability on nine-point scales compared to five (Yates, 2013). There is good internal consistency between the ratings given to the different domains of the mini-CEX, with a Cronbach's alpha of 0.79 (Weston & Smith, 2014).

It is noteworthy that in the 2012 update to the National Health Service (NHS) Foundation Programme curriculum for trainee doctors, the UK mini-CEX was updated to remove the tick boxes. Therefore the mini-CEX has moved away from using a GRS and instead focuses on

written feedback of strengths and weaknesses and the development of action plans (Weston & Smith, 2014). This was done in order to return the focus on the use of the mini-CEX as a formative tool. That said, the mini-CEX in its GRS form remains in use elsewhere.

Essential Skills Clusters (ESCs): A checklist approach

The standards for pre-registration nurses and midwives are set out by the Nursing & Midwifery Council (NMC) (2010). The standards incorporate, amongst other things, a set of mandatory *Essential Skills Clusters* (ESCs) which, according to Borneuf and Haigh (2010), developed out of concerns about skill deficits in earlier proficiency requirements. The standards state that the ESCs are to be used as guidance and should be incorporated into all pre-registration nursing and midwifery programmes, although the nature of programme incorporation is left to local determination (NMC, 2010).

ESCs encompass a broad set of interconnecting skills, knowledge and attitudes that are used to observe and assess trainee nurses and midwives. The ESCs comprise five skills clusters: care, compassion and communication; organisational aspects of care; infection prevention and control; nutrition and fluid management; and medicines management. Within these clusters there is a mixture of soft skills and knowledge content. For example, there are soft skills requirements to evidence that trainees "Form appropriate and constructive professional relationships with families and other carers" and "Manage and diffuse challenging situations effectively" (NMC, 2010, p.105 – *Care, compassion and communication ESC*). In other clusters there are requirements to demonstrate content knowledge such as "Recognises potential signs of infection and reports to relevant senior member of staff" (NMC, 2010, p.124 – *Infection prevention and control ESC*) and "Takes and records accurate measurements of weight, height, length, body mass index and other appropriate measures of nutritional status" (NMC, 2010, p.130 – *Nutrition and fluid management ESC*).

A variety of methods are used to assess ESCs, and these are characterised by a number of common elements. ESC assessment arrangements include:

- Mentor observation, with this usually organised around three meeting points (pre-, during-, and post-practice). This arrangement ensures that the assessment process performs both formative and summative functions
- The assessment materials articulate the criteria that are the basis for assessment
- Self-assessment is a key element of the assessment process. The assessment materials include space where the trainee is expected to record reflections on their practice and learning, and is in keeping with the tradition that reflection on practice has an important role in professional development, for example, Schön (1983)
- The assessment materials have an accountability function:
 - They are a record of attendance. This is because there are requirements that trainees complete a number of hours of practice that are attested to by the mentor.
 - They are a record of competence that is signed off by the mentor. The form of competence reporting for sign-off differs. Some ask the mentor to make a pass/fail judgement of

competence, others ask for a judgement of whether competent performance has been achieved in context(s), or ask for a judgement on the level of competence in terms of the trainee's participation involvement and the degrees of assistance required.

Comparing checklists and Global Rating Scales using systematic reviews

In this section we discuss accrued evidence about the advantages and disadvantages of checklists and GRSs. Results from many studies can be statistically combined in a systematic review, when studies meet particular quality criteria. Therefore, systematic reviews are useful for drawing evidence-based conclusions.

There are three systematic reviews which are key to our research topic. Ilgen et al. (2015) aimed to compare the reliability and validity of checklists and GRSs, as well as the correlation between scores from the two different scales. Their work was undertaken in the context of simulation-based assessment in health professionals' education. Their final analysis included 45 studies. McKinley et al. (2008) aimed to quantify the extent to which existing checklists allow for assessing both the humanistic and technical competencies needed in procedural competencies in the context of clinical procedures (tasks directly related to the care of a single patient, excluding physical examination). Their final analysis covered 75 studies. Finally, Ahmed, Miskovic, Darzi, Athanasiou, and Hanna (2011) aimed to identify assessment instruments and evaluate their validity and reliability in the context of direct observation of procedural/technical skills assessment in Medicine, (e.g., surgical skills). Such assessments may be work-based or simulations. Their final analysis included 106 studies.

The outcomes of an individual systematic review may not be generalizable to the assessment of complex competence across all professions. Furthermore, the outcome of the systematic reviews cannot be quantitatively combined or compared. Unfortunately, Ahmed et al. (2011) found that they could not statistically amalgamate results from different studies due to the diverse study designs. However, together, Ilgen et al. (2015) and McKinley et al. (2008) provide a solid evidence base from which to draw key comparisons between checklists and GRSs.

Reliability

Inter-rater reliability was substantial for both checklists and GRSs (Ilgen et al., 2015). Inter-item reliability was substantial for GRSs and lower for checklists. Interstation reliability was good for GRSs and suboptimal for checklists. Broadly speaking, the literature points towards GRSs achieving slightly better reliability than checklists.

Validity

Validation and development can be intense for task-specific checklists as each requires validation (Ilgen et al., 2015). In contrast, a GRS can be validated using evidence from many tasks yielding robust validity evidence, which can be less intense (Ilgen et al., 2015).

Ilgen et al. (2015) found that there was no difference between the content validity of checklists and GRSs. To evaluate content validity, researchers referred to previous instruments and expert consensus. On the other hand, McKinley et al. (2008) reviewed 88 checklists and

found that the inclusion of key competencies varied. The proportion of checklists including each competency was as follows:

Preparation: 74%,
Infection control: 32%,
Communication and working with the patient: 36%,
Teamworking: 15%,
Safety: 51%,
Procedural competence: 97%,
Post-procedural care: 27%.

Therefore, McKinley et al. (2008) argued that a GRS with a descriptor for each of these themes would have greater content validity than many checklists. Together, this information is a reminder that the quality of individual assessment instruments varies with several factors, including style of assessment (checklist or GRS).

Regarding criterion validity, Ilgen et al. (2015) found that the criterion validity was equivalent for checklists and GRSs in 11 studies, and higher for GRSs in a further 6 studies. Furthermore, Ilgen et al. (2015) reported there was a correlation of 0.76 between checklist and GRS measures, denoting that they measured somewhat similar traits. On balance, checklists and GRSs may measure similar traits, but GRSs generally have higher criterion validity.

The outcomes of rater training were under reported (Ilgen et al., 2015). To be specific, one study about checklists and two about GRSs reported rater training outcomes. This resonates with the point made earlier that there is little research about rater training for the mini-CEX. Therefore, rater training is likely to be an area requiring further research.

The scope of systematic reviews

There are several factors which were not included in either systematic review. These included cognitive validity (whether the raters or test takers used the intended cognitive activities). An example of a single study that addresses cognitive validity is McIlroy, Hodges, McNaughton, and Regehr (2002). They found that students adapt their behaviours according to their perceptions and expectations of the measurement tool being used to assess them. A total of 57 medical students assigned to 2 groups were primed to expect that they were being assessed on a 10-station Objective Structured Clinical Examination (OSCE) with either a GRS or checklist measure. McIlroy et al. (2002) found a significant interaction between the type of OSCE measure and the measure students expected to be used. Those in the group anticipating a checklist attained higher checklist scores but lower GRS scores than those in the group anticipating a GRS assessment, although the effect size was small. They also found higher interstation reliability coefficients for the GRS ratings than for the checklist scores across all students, thus suggesting that overall GRS ratings show higher reliability regardless of students' perceptions. The difference in interstation reliability between the two measures was greater for the students expecting to be assessed using a GRS, which also showed lower interstation reliability for both measures. The researchers speculated that when students expect to be assessed using GRSs, their performance became more heterogeneous across stations. This may be because the students are less able to rely on a 'script' and so their performance varies according to their content-

specific expertise on each station, thus decreasing reliability. This study shows that the remit of the systematic reviews is somewhat limited. Nonetheless, the systematic reviews provide rich and solid evidence regarding many validity and reliability issues. Together the systematic reviews reveal that GRSs tend to achieve greater validity than checklists.

Conclusions

The aim of our research was to explore the measurement qualities of checklists and GRSs in the context of assessing complex competence. Firstly, we reviewed the literature about the affordances of human judgement and mechanical combination of human judgements. Secondly, we considered examples of assessment instruments (checklists and GRSs) used to assess complex competence in highly regarded professions. These examples served to contextualise and illuminate assessment issues. Finally, we compiled research evidence from the outcomes of systematic reviews which compared advantages and disadvantages of checklists and GRSs. Our research has caveats: for example, focusing on healthcare may restrict the generalizability of the findings. However, merging the research on human judgement, mini-CEX, essential skills and systematic reviews provides a nuanced and firm evidence base for drawing key conclusions.

Reliability

The weight of evidence signifies that GRSs generally achieve better reliability than checklists. Furthermore, human judgement research tends to confirm that accuracy is enhanced by humans judging individual traits and those judgements being mechanically combined to gain an overall assessment. Technology in this area is ever advancing, including deriving sentiment levels from assessors' comments and combining them with other quantitative data to report assessment outcomes (Janssen et al., 2016). Hence, we recommend that human judgements focus on judging individual traits and that these judgements are combined by computer, when practicable.

Validity

Together the systematic reviews suggest that GRSs tend to achieve greater validity than checklists. However, validity is a multifaceted concept and the picture is nuanced. There is no difference between the content validity of checklists and GRSs; however the content validity of individual instruments varies. Whilst checklists and GRSs can measure similar traits, the criterion validity of GRSs is generally slightly higher. In summary, it is recommended that GRSs are considered preferable to checklists, although a high-quality checklist is better than a poor-quality GRS.

Social bias

Concerns about assessor bias are a common feature to both checklist and GRS approaches. For example, studies of mini-CEX show that rating scores appear to be influenced by the nature of the assessor, such as whether they were a resident or a faculty member (Al Ansari et al., 2013). Social biases can also extend to the contextual features that surround an assessment. In the case of the mini-CEX, evidence

suggests that assessing clinical competence in a real-life setting may result in more lenient judgements compared with simulated task environments (Hawkins et al., 2010). For ESC assessment there are concerns that assessors' dual practice and assessment roles can interfere with the assessment process as maintaining interpersonal relations can potentially influence assessor judgements (Heaslip & Scammell, 2012). This has parallels with findings in other vocational areas, for example, Colley and Jarvis (2007) and Yaphe and Street (2003). Broadly speaking there are three ways of guarding against social bias: rater training; moderation; and scaling. It is recommended that such safeguards are employed.

Practicalities

The practicability of assessment is also a feature that influences the use of both checklist and GRS assessment approaches. In general it is considered that holistic judgements can work well in contexts that afford frequent and close observations of learner performance, for example, Curtis (2004). At the same time, contextual considerations can undermine the enactment of multiple assessment observations. Assessment in professional contexts can be resource intensive. For example, it is suggested that the validity of the mini-CEX requires different assessors to assess a range of clinical skills over time in a number of contexts and scenarios, and that this should involve 10 encounters and between 6 and 10 different assessors (Norcini et al., 2003). Similarly, the assessment of ESCs often involves mentor observations that are organised around three meeting points during a placement. Evidence suggests that it is sometimes difficult to ensure that devolved mentor assessment responsibilities are carried out at the appropriate time during the placement (Shaw, 2016). This issue has led, in part, to the development of e-portfolio tools to support the assessment process. Such systems often also facilitate combining human judgements on multiple traits and assessments, as we have mentioned. It is recommended that those making assessment judgements are involved in designing assessments to increase the manageability of the assessments.

Evidence quality

The validity of using observation as an assessment tool links to the notion that the method elicits characteristics of performance that are indicative of 'true' capability. To support this, it has been noted that past (observed) performance can be taken as a good indicator of future performance – see Adams (2012), cited in Marks (2014). The quality of an assessment relates to the quality of the evidence that is elicited through an assessment task. Therefore any justification for using observation as an assessment tool relates to the quality of the instruments that support that observation process. For both GRS and checklist approaches, there are variances around the practices that are found in different contexts, and this can undermine the confidence of assessment outcomes. For example, the reliability of the mini-CEX assessment is influenced by the number of points on the rating scale (Yates, 2013). In the case of ESC assessment, it is noted that the form of competence reporting for sign-off can differ but that the role of competent professionals (i.e., mentors) is generally limited to a sign-off function that attests to task completion rather than the quality of performance. This highlights the importance of including validation in the development and review processes.

References

- Adams, R. (2012). *National Partnership Agreement on Literacy and Numeracy reporting: Measures and models for reporting gain over time*. Sydney, Australian: Council of Australian Governments Reform Council.
- Ahmed, K., Miskovic, D., Darzi, A., Athanasiou, T., & Hanna, G. B. (2011). Observational tools for assessment of procedural skills: a systematic review. *The American Journal of Surgery*, 202, 469–480. Available online from doi: 10.1016/j.amjsurg.2010.10.020
- Al Ansari, A., Ali, S. K., & Donnon, T. (2013). The construct and criterion validity of the mini-CEX: a meta-analysis of the published research. *Academic Medicine*, 88(3), 413–420. Available online from doi: 10.1097/ACM.0b013e318280a953
- Boritz, J. E., & Timoshenko, L., M. (2014). On the Use of Checklists in Auditing: A commentary. *Current Issues in Auditing*, 8(1), C1–C25. Available online from doi: 10.2308/ciia-50741
- Borneuf, A. M., & Haigh, C. (2010). The who and where of clinical skills teaching: A review from the UK perspective. *Nurse Education Today*, 30(2), 197–201. Available online from doi: 10.1016/j.nedt.2009.07.012
- Boursicot, K., Etheridge, L., Setna, Z., Sturrock, A., Ker, J., Smees, S., & Sambandam, E. (2011). Performance in assessment: Consensus statement and recommendations from the Ottawa conference. *Medical Teacher*, 33(5), 370–383. Available online from doi: 10.3109/0142159X.2011.565831
- Colley, H., & Jarvis, J. (2007). Formality and informality in the summative assessment of motor vehicle apprentices: a case study. *Assessment in Education: Principles, Policy & Practice*, 14(3), 295–314. Available online from doi: 10.1080/09695940701591883
- Curtis, D. D. (2004). The assessment of generic skills. In J. Gibb (Ed.), *Generic skills in vocational education and training: research findings* (pp.136–156). Station Arcade, Australia: National Centre for Vocational Education Research.
- Domingues, R. C. L., Amaral, E., & Zeferino, A. M. B. (2009). Global overall rating for assessing clinical competence: what does it really show? *Medical Education*, 43(9), 883–886. Available online from doi: 10.1111/j.1365-2923.2009.03431.x
- Dreyfus, S. E., & Dreyfus, H., L. (1980). *A Five-Stage Model of the Mental Activities Involved in Direct Skill Acquisition*: Operations Research Center, University of California.
- Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organizational Behaviour and Human Performance*, 7, 86–106. Available online from doi: 10.1016/0030-5073(72)90009-8
- Eining, M., Jones, D. R., & Loebbecke, J. (1997). Reliance on decision aids: An examination of auditors' assessment of management fraud. *Auditing*, 16(2), 1–19.
- Eraut, M. (1994). *Developing Professional Knowledge and Competence*. London: Falmer Press.
- Eraut, M., & Steadman, S. (1998). *Evaluation of Level 5 Management S/NVQs*. Brighton: University of Sussex Institute of Education.
- Feltham, R. (1988). Assessment centre decision making: judgement vs mechanical. *Journal of Occupational Psychology*, 61, 237–241.
- Gilovich, T., & Griffin, D. (2002). Introduction – heuristics and biases: then and now. In T. Gilovich, D. Griffin, & D. Kahnemann (Eds.), *Heuristics and biases: the psychology of intuitive judgement*. (pp.1 to 18). Cambridge: Cambridge University Press.
- Gilovich, T., Griffin, D., & Kahnemann, D. (2002). *Heuristics and biases: the psychology of intuitive judgement*. Cambridge: Cambridge University Press.
- Gingerich, A., van der Vleuten, C. P., Eva, K. W., & Regehr, G. (2014). More consensus than idiosyncrasy: Categorizing social judgments to examine variability in Mini-CEX ratings. *Academic Medicine*, 89(11), 1510–1519. Available online from doi: 10.1097/ACM.0000000000000486

- Hawkins, R. E., Margolis, M. J., Durning, S. J., & Norcini, J. J. (2010). Constructing a validity argument for the mini-Clinical Evaluation Exercise: a review of the research. *Academic Medicine*, 85(9), 1453–1461. Available online from doi: 10.1097/ACM.0b013e3181eac3e6
- Heaslip, V., & Scammell, J. M. (2012). Failing underperforming students: The role of grading in practice assessment. *Nurse Education in Practice*, 12(2), 95–100. Available online from doi: 10.1016/j.nepr.2011.08.003
- Highhouse, S., & Kostek, J., A. (2013). Holistic assessment for selection and placement. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA Handbook of Testing and Assessment in Psychology* (Vol. 1. Test theory and Testing and Assessment in Industrial and Organizational Psychology). Washington, DC, US: American Psychological Association.
- Ilgen, J. S., Ma, I. W. Y., Hatala, R., & Cook, D. A. (2015). A systematic review of validity evidence for checklists versus global rating scales in simulation based assessment. *Medical Education in Review*, 49, 161–173. Available online from doi: 10.1111/medu.12621
- Janssen, D., Holthuijsen, M., Clarebout, G., Donkers, J., Slof, B., & van der Schaaf, M. (2016). *Learning Analytics enhanced E-portfolios for Workplace Based Assessment*. Paper presented at the European Association for Research on Learning and Instruction (EARLI) Conference, SIG 1 Assessment and Evaluation, Universität München, Munich.
- Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Doubleday.
- Laming, D. (2004). *Human judgment: the eye of the beholder*. Hong Kong: Thomson Learning.
- Lester, S. (2000). The Professional Accreditation of Conservator-Restorers: Developing a competence-based professional assessment system. *Assessment & Evaluation in Higher Education*, 25(4), 407–419. Available online from doi: 10.1080/713611439
- Marks, G. N. (2014). Demographic and socioeconomic inequalities in student achievement over the school career. *Australian Journal of Education*, 58(3), 223–247. Available online from doi: 10.1177/0004944114537052
- McIlroy, J. H., Hodges, B., McNaughton, N., & Regehr, G. (2002). The effect of candidates' perceptions of the evaluation method on reliability of checklist and global rating scores in an objective structured clinical examination. *Journal of Medical Education*, 77(7), 725–728. Available online from doi: 10.1097/00001888-200207000-00018
- McKinley, R. K., Strand, J., Ward, L., Gray, T., Lun-Jones, T., & Miller, H. (2008). Checklists for assessment and certification of clinical procedural skills omit essential competencies: a systematic review. *Medical Education*, 42, 338–349. Available online from doi: 10.1111/j.1365-2923.2007.02970.x
- Norcini, J. J. (2005). The Mini Clinical Evaluation Exercise (mini-CEX). *The Clinical Teacher*, 2(1), 25–30. Available online from doi: 10.1111/j.1743-498X.2005.00060.x
- Norcini, J. J., Blank, L. L., Duffy, F. D., & Fortna, G. S. (2003). The Mini-CEX: A method for assessing clinical skills. *Annals of Internal Medicine*, 138(6), 476–481. Available online from doi: 10.7326/0003-4819-138-6-200303180-00012
- Nursing & Midwifery Council. (2010). *Standards for Pre-Registration Nursing Education*. London: NMC.
- Pelgrim, E. A. M. (2010). In-training assessment using direct observation of single-patient encounters: a literature review. *Advances in Health Sciences Education*, 16(1), 131–142. Available online from doi: 10.1007/s10459-010-9235-6
- Pulito, A. R., Donnelly, M. B., & Plymale, M. (2007). Factors in faculty evaluation of medical students' performance. *Medical Education*, 41(7), 667–675. Available online from doi: 10.1111/j.1365-2923.2007.02787.x
- Schön, D. (1983). *The Reflective Practitioner: How professionals think in action*. New York: Basic Books.
- Shaw, S. (2016). *Go-Electronic practice assessment in action*. Paper presented at the Electronic Practice Assessment – Nurses and Midwives Conference, Anglia Ruskin University, Cambridge. Retrieved from https://www.youtube.com/watch?v=ji8EFfyDms&list=PLI7A9-faRI96oce_7qR5pefrqeWNjXk7M&index=4
- Tversky, A., & Kahneman, D. (1982). Judgement under uncertainty: Heuristics and biases. In D. Kahneman, P. Slovic, & T. A. (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp.3–22). Cambridge: Cambridge University Press.
- Watson, A. (1994). Strategies for the Assessment of Competence. *The Vocational Aspect of Education*, 46(2), 155–165. Available online from doi: 10.1080/0305787940460205
- Weston, P. S., & Smith, C. A. (2014). The use of mini-CEX in UK foundation training six years following its introduction: lessons still to be learned and the benefit of formal teaching regarding its utility. *Medical Teacher*, 36(2), 155–163. Available online from doi: 10.3109/0142159X.2013.836267
- Winter, R. (1995). The assessment of professional competences: the importance of general criteria. In A. Edwards & P. Knight (Eds.), *Assessing Competence in Higher Education*. London: Kogan Page.
- Yaphe, J., & Street, S. (2003). How do examiners decide? A qualitative study of the process of decision making in the oral component of the MRCPG examination. *Medical Education*, 37(9), 764–771. Available online from doi: 10.1046/j.1365-2923.2003.01606.x
- Yates, P. J. (2013). The Mini-CEX is not Valid or Reliable in Assessing the Clinical Competence of Higher Surgical Trainees. *The Bulletin of the Royal College of Surgeons of England*, 96(8), 1–4. Available online from doi: 10.1308/147363513X13690603820144

Statistics Reports

The Research Division

Examinations generate large volumes of statistical data (approximately 800,000 candidates sit general qualifications each year in the UK). The on-going *Statistics Reports Series* provides statistical summaries of various aspects of the English examination system. The objective of the series is to provide statistical information about the system, such as trends in pupil uptake and attainment, qualifications choice, subject combinations and subject provision at school. The reports, mainly produced using national-level examination data, are available in both PDF and Excel format on our website:

www.cambridgeassessment.org.uk/our-research/all-published-resources/statistical-reports/

The most recent additions to the series are:

- *Statistics Report Series No.111: Candidates awarded A* and A grades at GCSE in 2015*
- *Statistics Report Series No.112: Ranking of candidates' best GCSE grades by subject in 2015.*



Cambridge Assessment
Network

Assessing the world – visiting Cleverlands

A seminar with Lucy Crehan, author of 'Cleverlands: the secrets behind the success of the world's education superpowers'

12 October 2017 | 16.00 – 18.30 | Cambridge | Free

Register today: www.canetwork.org.uk/seminars

Lucy Crehan, former teacher and international education consultant will be sharing stories and research findings from her travels to the 'top-performing' education systems in six countries on four continents at this Cambridge Assessment Network seminar.

In documenting her teaching odyssey across the globe, Lucy has linked her experiences to key strands in educational theory and research, giving deep insights into both the culture and practices in a range of key jurisdictions and illuminating major educational discussions of curriculum and assessment.

Research News

Karen Barden Research Division

Conferences and seminars

Royal Statistical Society (RSS)

The RSS 2016 International Conference took place at the University of Manchester in September 2016. Now in its 24th year, the RSS conference welcomes all statisticians and users of data, providing a crucial platform for the discussion and debate of statistical topics. Ellie Darlington and Jessica Bowyer, Research Division, took part in the session *Communicating Statistics: Statistical education in schools*. They presented a paper on *Statistical education at A level: an international perspective*.

British Educational Research Association (BERA)

The University of Leeds hosted the BERA Annual Conference in September 2016. This provided an opportunity to develop new research ideas, and to build new research relationships within the research education community. Several researchers from Cambridge Assessment attended the conference and the following papers were presented:

Carmen Vidal Rodeiro, Research Division: *The study of Modern Foreign Languages in England: uptake in secondary school and progression to Higher Education*.

Tim Gill, Research Division: *An analysis of the effect of taking the EPQ on performance in other level 3 qualifications*.

Lorna Stabler, Cambridge International Examinations: *Validating an Art and Design qualification: evidence for the validity of a performance-based assessment*.

Association for Educational Assessment-Europe (AEA-Europe)

The 17th AEA-Europe Annual Conference took place in Limassol, Cyprus in November 2016 with the theme *Social and political underpinnings of educational assessment: Past, present and future*. Several researchers from Cambridge Assessment attended the conference and the following papers were presented:

Tom Bramley, Research Division: *Investigating experts' perceptions of examination question demand*.

Tom Benton, Research Division: *Evidence for the reliability of coursework*.

Victoria Crisp, Martin Johnson and Filio Constantinou, Research Division: *'Question quality': The concept of quality in the context of exam questions*.

Simon Child, OCR, and Stuart Shaw, Cambridge International Examinations: *Utilising technology in the assessment of collaboration: a critique of PISA's collaborative problem solving tasks*.

Simon Child and Martina Kuvalja, OCR: *What makes a good seeding script? Perceptions from Principal Examiners of a UK awarding body*.

Martin Johnson, Filio Constantinou, and Victoria Crisp, Research Division: *How do question writers compose examination questions? Question writing as a socio-cognitive process*.

Tim Oates, Assessment Research and Development, and Sylvia Green, Research Division: *Shifting emphases: qualifications, accountability and school improvement*.

Nadir Zanini, Research Division: *Do tiered examinations affect candidates' achievement? Some empirical evidence on Modern Foreign Languages*.

The following poster was also presented:

Sarah Hughes, Stuart Shaw, Lorna Stabler, and Magda Werno, Cambridge International Examinations: *Does the mode of standardisation matter? The effect on reliability of marking and marker perceptions*.

Further details about the AEA-Europe conference presentations can be found on our website: www.cambridgeassessment.org.uk/events/aea-annual-conference-2016/

Further information on all conference papers can be found on our website: www.cambridgeassessment.org.uk/our-research/all-published-resources/conference-papers/

The Cambridge Approach to Textbooks

Tim Oates, CBE, Group Director of Assessment Research and Development, launched *The Cambridge Approach to Textbooks* in April 2016¹. This set of criteria is in direct response to Tim's identification of "England's need to restore the primacy of 'real' textbooks worldwide" and follows on from *The Cambridge Approach to Assessment* published in 2009².

Tim presented *The Cambridge Approach to Textbooks* live from Westminster, London and was joined by leading experts in the field to help detail how the role of textbooks and allied learning resources has developed in the UK and internationally.

The seminar also heard from Professor David Lambert, University College London, and Lord Knight, Chief Education Adviser, TES Global. Other speakers contributing to the event included Fei Chen Lee, Times Publishing Singapore; Debbie Morgan, National Centre for Excellence in the Teaching of Mathematics; Lis Tribe, Hodder Education Group; Bron Duly, RM Books; and Jane Mann, Director of Education Reform, Cambridge University Press.

The presentations were accompanied by discussion and debate with the attending senior education experts, and online via a live-streamed video. Further details of the event, video highlights, and a range of related materials can be found on our website: www.cambridgeassessment.org.uk/news/launch-of-the-cambridge-approach-to-textbooks/

1. Oates, T. (2016). *The Cambridge Approach. Principles for designing high-quality textbooks and resource materials*. Cambridge, UK: Cambridge Assessment. Available online at: <http://www.cambridgeassessment.org.uk/Images/299335-the-cambridge-approach-to-textbooks.pdf>

2. Oates, T. (2009). *The Cambridge Approach. Principles for designing, administering and evaluating assessment*. Cambridge, UK: Cambridge Assessment. Available online at: <http://www.cambridgeassessment.org.uk/Images/109848-cambridge-approach.pdf>

Aspects of Writing

A Cambridge Assessment research seminar took place in London in November 2016 to launch the report of the 2014 phase of *Variations in Aspects of Writing in 16+ English Examinations*. The report, written by Gill Elliott, Sylvia Green, Filio Constantinou, Sylvia Vitello, Lucy Chambers, Nicky Rushton, Jo Ireland, Jessica Bowyer and David Beauchamp, has triggered great debate across the educational research and English teaching communities, and the media.

The report is the latest phase of a unique study, which has been carried out every 10 years since the 1980s, and explores changes in a range of aspects of students' writing in the context of formal English exams between 1980 and 2014. The aspects of writing under scrutiny include spelling, punctuation, sentence structure and the use of paragraphs. Access to a rich corpus of writing from recent decades has afforded Cambridge Assessment researchers an invaluable opportunity to conduct this cross-sectional study which provides insights that will interest researchers, teachers and the educational community.

Professor Debra Myhill, Professor of Education and Pro-Vice-Chancellor, University of Exeter, set the scene and provided the context for the study. Professor of English and Linguistics Director at Aston University, Professor Ursula Clark, explored *Teaching grammar: where do we go from here?*

Presentations on various aspects of the research were made by researchers from the Research Division: Sylvia Green set out the background to the study; Gill Elliott talked through the method of the 2014 phase of the study and put into context the findings set out in the report; Filio Constantinou looked at the impact the ever increasing use of social media may or may not have had on students' writing; and Nicky Rushton addressed the change in common misspellings by students between 1980 and 2014.

The full report was published as *Research Matters* Special Issue 4 and is available from our website: www.cambridgeassessment.org.uk/Images/340982-research-matters-special-issue-4-aspects-of-writing-1980-2014.pdf If you would like to receive a printed copy, please email your contact details to researchprogrammes@cambridgeassessment.org.uk

Further details of the seminar, video highlights, audience views and a range of additional resources can be found on our website: www.cambridgeassessment.org.uk/aspects-of-writing/

Publications

The following articles have been published since *Research Matters*, Issue 22:

- Bowyer, J. and Darlington, E. (2017). Mathematical struggles and ensuring success: post-compulsory mathematics as preparation for undergraduate bioscience. *Journal of Biological Education*. Advance online publication available at: <http://www.tandfonline.com/doi/full/10.1080/00219266.2017.1285803>
- Bowyer, J. and Darlington, E. (2017). Should I take Further Mathematics? Physics undergraduates' experiences of post-compulsory Mathematics. *Physics Education*, 52(1). Advance online publication available at: <http://iopscience.iop.org/article/10.1088/1361-6552/52/1/015007>

- Bramley, T. and Crisp, V. (2017). Spoilt for choice? Issues around the use and comparability of optional exam questions. *Assessment in Education: Principles, Policy & Practice*. Advance online publication available at: <http://www.tandfonline.com/doi/full/10.1080/0969594X.2017.1287662>
- Crisp, V. (2017). The judgement processes involved in the moderation of teacher-assessed projects. *Oxford Review of Education*, 43(1), 19–37. Available online at: <http://www.tandfonline.com/doi/full/10.1080/03054985.2016.1232245>
- Darlington, E. and Bowyer, J. (2016). Engineering undergraduates' views of A-level Mathematics and Further Mathematics as preparation for their degree. *Teaching Mathematics and its Applications*. Advance online publication available at: <https://doi.org/10.1093/teamat/hrw020>
- Darlington, E. and Bowyer, J. (2016). Students' views of, and motivations for, studying A-level Further Mathematics. *MSOR Connections*, 15(1), 4–13. Available online at: <https://journals.gre.ac.uk/index.php/msor/issue/view/issue/47/37>
- Darlington, E. and Bowyer, J. (2016). How well does A-level Mathematics prepare students for the mathematical demand of chemistry degrees? *Chemistry Education Research and Practice*. Available online at: <http://pubs.rsc.org/en/content/articlelanding/2016/RP/C6RP00170J#divAbstract>
- Gill, T. (2017). Preparing students for university study: a statistical comparison of different post-16 qualifications. *Research Papers in Education*. Advance online publication available at: <http://dx.doi.org/10.1080/02671522.2017.1302498>
- Johnson, M. (2016). The challenges of researching digital technology use: examples from an assessment context. *International Journal of e-Assessment*, 1(2), 1–10.
- Johnson, M. and Oates, T. (2016). Making sense of a learning space: How freestyle scooter-riders learn in a skate park. *Informal Learning Review*, 140, 17–21. Available online at: <http://www.informallearning.com/the-informal-learning-review.html>
- Shaw, S. and Werno, A. (2016). Preparing for college success: exploring the impact of the High School Cambridge Acceleration Program on US university students. *College and University: Educating the Modern Higher Education Administration Professional*, 91(4), 2–21.
- Wilson, F., Child, S., and Suto, I. (2016). Assessing the transition between school and university: Differences in assessment between A level and university in English. *Arts and Humanities in Higher Education*, 41(1), 1–21. Available online at: <http://journals.sagepub.com/doi/pdf/10.1177/1474022216628302>

Further information on all journal papers and book chapters can be found on our website: www.cambridgeassessment.org.uk/our-research/all-published-resources/journal-papers-and-book-chapters/

Reports of research carried out by the Research Division for Cambridge Assessment and our exam boards, or externally funded research carried out for third parties, including the regulators in the UK and many ministries overseas, are also available from our website: www.cambridgeassessment.org.uk/our-research/all-published-resources/research-reports/

Data Bytes

The Data & Analytics Team Research Division


Data Bytes is a series of data graphics from the Research Division, designed to bring the latest trends and research in educational assessment to a wide audience. Each Data Byte consists of a single graphic or interactive visualization designed to present a notable data set or research finding relevant to educational assessment. The graphic is accompanied by a brief text explaining what the image shows and why it is significant. Topics are often chosen to coincide with contemporary news or recent Cambridge Assessment research outputs.

Since the series launched in October 2015 we have published the following *Data Bytes*, all of which can be found at www.cambridgeassessment.org.uk/our-research/data-bytes/ Interactive graphics are marked with (I).

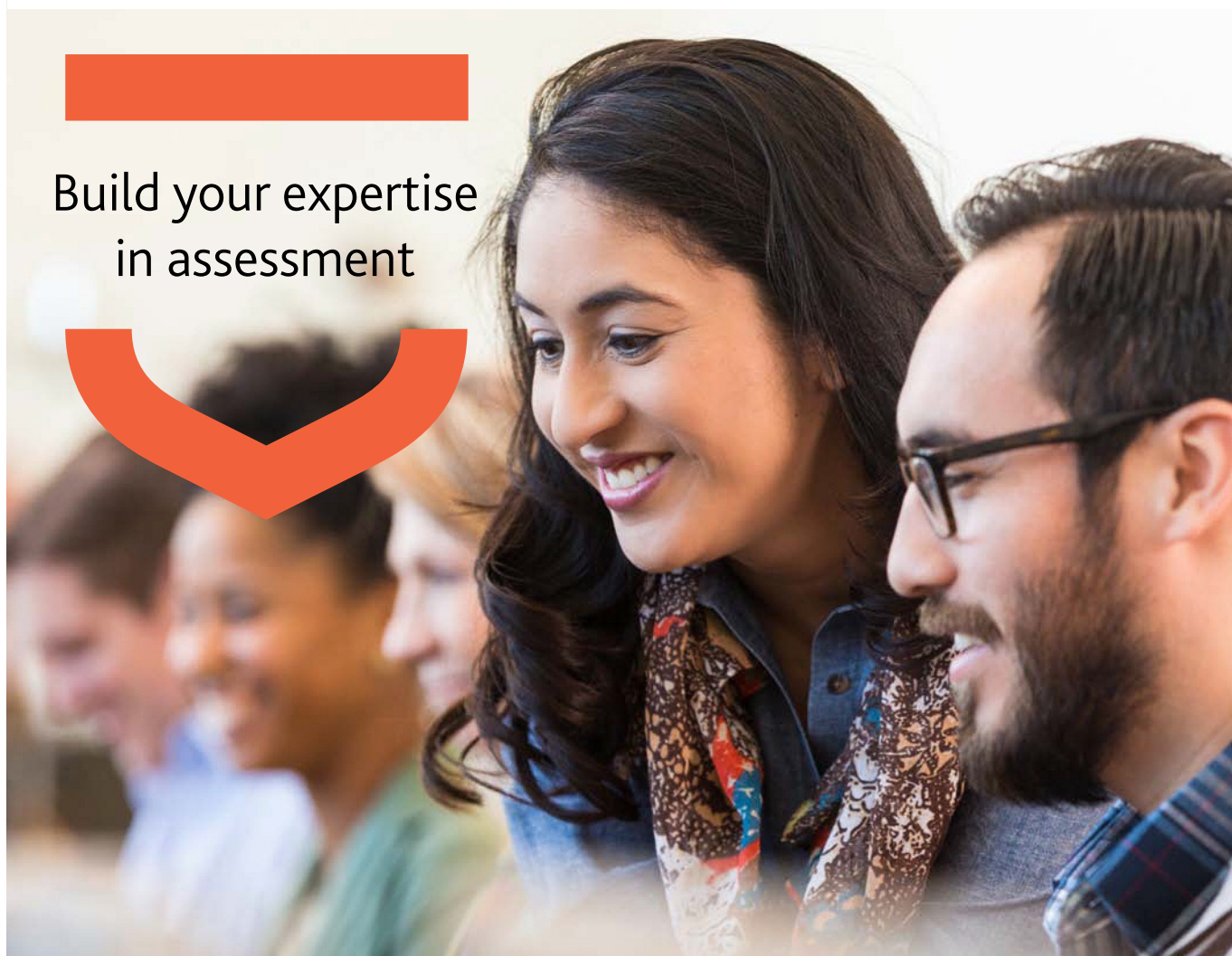

- October 2015: *The gender gap in attainment at GCSE*
- November 2015: *The changing gender gap*
- December 2015: *The effect of A* as a predictor of university performance* (I)
- January 2016: *Most popular A level subjects since 2000*
- February 2016: *Global trends in primary, secondary and post-secondary educational attainment*
- March 2016: *The role of teaching styles in Mathematics achievement*
- April 2016: *Teacher mobility within the EU*
- May 2016: *Key statistics on the Extended Project Qualification 2008–2015*
- July 2016: *The average age of teachers in secondary schools*
- August 2016: *Tweeting about exams*
- October 2016: *What GCSE and A Level subject combinations are offered by examination centres in England?* (I)
- November 2016: *Student destinations from Key Stage 5*
- December 2016: *Candidates' best GCSE grades*
- January 2017: *Re-sit rates for A Level subjects*
- February 2017: *The international popularity of STEM subjects*
- March 2017: *Popularity of Level 3 vocational subjects.*



Cambridge Assessment Network



Build your expertise
in assessment



Cambridge Assessment Network provides professional development for the assessment community in the UK and internationally.

We equip education professionals with the tools, knowledge and understanding to be confident and capable assessment practitioners.

See our
training and events
programme for 2017
www.canetwork.org.uk

A new look for *Research Matters*

Karen Barden Research Division

You will have noticed a new look to this issue of *Research Matters* to coincide with the launch of the new Cambridge Assessment brand. Gone is the old 'A' logo and 'swoosh' and in its place is a return to the University of Cambridge coat of arms. The new logo allows us to highlight our unique position as the oldest exams group still in existence and the only one still attached to a university. It also aligns our work more closely with that of the University and other members of the University family with which we work including Cambridge University Press, the Faculties of Education and Mathematics and various other departments. Our new brand reflects both how we have grown as an

international organisation and how the world has changed since we became Cambridge Assessment in 2005 and we published the very first issue of *Research Matters*.

Further details of the launch of the new Cambridge Assessment brand and how we are evolving to support our customers can be found on our website: www.cambridgeassessment.org.uk/news/evolving-to-support-our-customers-launching-the-new-cambridge-assessment-brand/

Research Matters will now be published in spring and autumn each year. All issues published since 2005 are available from our website: www.cambridgeassessment.org.uk/research-matters

Contents / Issue 23 / Spring 2017

- 2 Tweeting about exams: Investigating the use of social media over the summer 2016 session : Tom Sutch and Nicole Klir
- 10 The clue in the dot of the 'i': Experiments in quick methods for verifying identity via handwriting : Tom Benton
- 17 Evaluating blended learning: Bringing the elements together : Jessica Bowyer and Lucy Chambers
- 27 An analysis of the effect of taking the EPQ on performance in other Level 3 qualifications : Tim Gill
- 35 A review of instruments for assessing complex vocational competence : Jackie Creatorex, Martin Johnson and Victoria Coleman
- 43 Statistics Reports : The Research Division
- 44 Research News : Karen Barden
- 46 Data Bytes : The Data & Analytics Team
- 48 A new look for *Research Matters* : Karen Barden

Cambridge Assessment

1 Hills Road
Cambridge CB1 2EU
United Kingdom

+44(0)1223 552666
researchprogrammes@cambridgeassessment.org.uk
www.cambridgeassessment.org.uk

© UCLES 2017



ISSN: 1755–6031