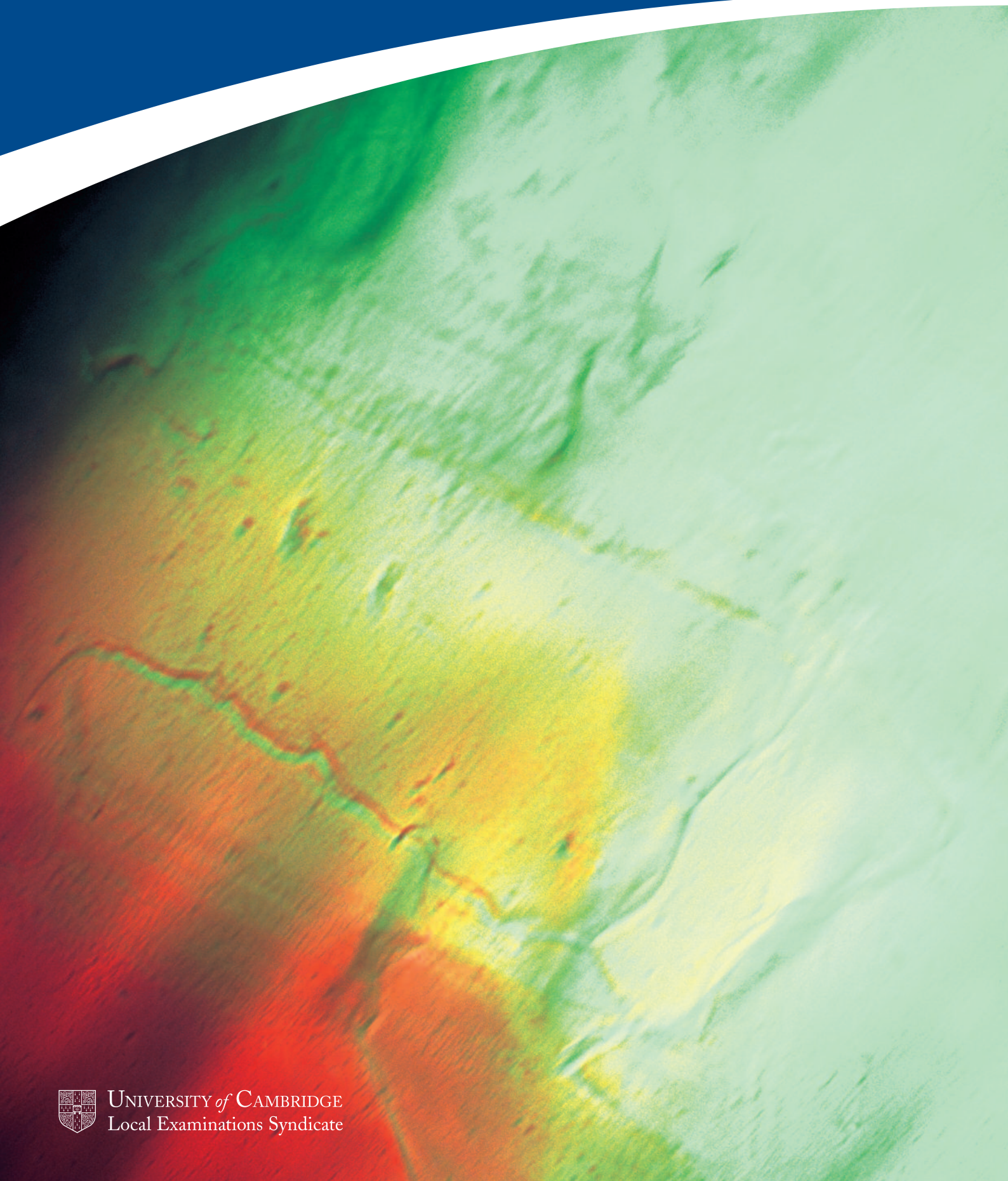
Issue 19 Winter 2015

Research Matters

CAMBRIDGE ASSESSMENT

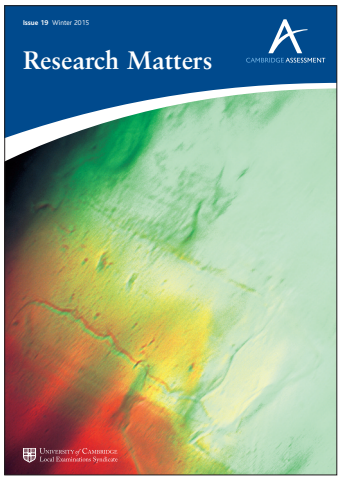


University of Cambridge
Local Examinations Syndicate

CAMBRIDGE ASSESSMENT

Citation

Articles in this publication should be cited as: Child, S.F.J., Darlington, E., and Gill, T. (2014). A level History choices: Which factors motivate teachers' unit and topic choices? *Research Matters: A Cambridge Assessment Publication*, *19*, 2–6.



1 **Foreword** : Tim Oates

1 **Editorial** : Sylvia Green

2 **A level History choices: Which factors motivate teachers' unit and topic choices?** : Simon Child, Ellie Darlington and Tim Gill

7 **Context led Science courses: A review** : Frances Wilson, Steve Evans and Sarah Old

14 **Assessing active citizenship: An international perspective** : Prerna Carroll, Simon Child and Ellie Darlington

19 **An investigation into the numbers and characteristics of candidates with incomplete entries at AS/A level** : Carmen Vidal Rodeiro

26 **The moderation of coursework and controlled assessment: A summary** : Tim Gill

31 **Reflections on a framework for validation - Five years on** : Stuart Shaw and Victoria Crisp

38 **Text Mining: An introduction to theory and some applications** : Nadir Zanini and Vikas Dhawan

45 **Research News** : Karen Barden

47 **Statistical Reports** : Tim Gill

47 **Cambridge Mathematics launch**

48 **Stop Press: CBE for Tim Oates**

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 19* and all previous issues are available from our website: www.cambridgeassessment.org.uk/our-research

Research Matters : 19

A CAMBRIDGE ASSESSMENT PUBLICATION

Foreword

As this issue of *Research Matters* goes to press, two important matters remain highly controversial issues in qualifications policy in England – the move in Science from assessment of practical work, and the 'de-coupling' of AS and A level examinations – to a position where AS level will continue to exist, but performance in the AS level will not contribute to the grade in the corresponding A level. Practical work in Science is vital, for developing an awareness of handling of materials and equipment, for encouraging deep learning through engaging activities, and for developing competence in design and control, observation and reporting. But, in specific qualifications, by 2013 we had moved to a position where we lacked clarity in its precise purpose – as my own and Robin Millar's work has highlighted – and we have pursued highly dependable assessment at the expense of rich learning. The compromise position adopted by Ofqual aims to re-set this relation and introduce clarity into purpose. As a nation we now need to monitor closely the impact, on learning programmes and on attainment, of the revised assessment requirements. The sense of controversy around the new qualifications will only subdue when evidence of a retention of high-quality practical work emerges from schools and colleges. Likewise with AS level; prominent voices continue to be heard on both sides of the AS debate. The analysis of 'four AS followed by three A levels', presented in this issue, highlights the interesting benefits of being able to refine subject choice by discontinuing study in a subject, at the end of the first year of advanced study. This does not mean that AS needs to contribute to the final A2 grade; but – as with Science coursework – assessment-dominated thinking has led many schools to move from four AS to three A levels, despite the advantages of (i) gaining individual, social and economic benefit by helping students focus on subjects which they enjoy and/or in which they excel; and (ii) providing Higher Education institutions with dependable information at the end of Year 12, which enhances Higher Education offer-making. The approach in Science need not reduce the amount of engaging practical work – but it may so do. The de-coupling of AS level need not reduce the numbers using AS to refine their choices – but it may so do. The proof of the pudding will be in the eating.

Tim Oates *Group Director, Assessment Research and Development*

Editorial

The importance of a sound research evidence base to underpin qualifications reform is reflected in the articles in this issue. The first is an extension of the work conducted by Child, Darlington and Gill reported in Issue 18 which explored the choices of topics and units made by students and teachers in A level History. Their more recent research examines the factors that influence those choices and analyses the motivations underlying those decisions. Wilson, Evans and Old focus on Science and the need for sustained growth in uptake following a recent increase in the percentage of A level entries for Science. As a result of concerns about students' abilities in applying scientific concepts, they examine a context led approach to Science courses which have been developed in an attempt to address these concerns. Assessment of Citizenship is another challenging area in qualifications development. Carroll, Child and Darlington discuss the assessment of GCSE Citizenship. They explore definitions of Citizenship, international approaches to its assessment, and different approaches to external examination of the subject.

The next two articles address more technical aspects of reforms to qualifications. Vidal Rodeiro's research aims to gain an understanding of the numbers and types of students who start but do not complete their AS and A level qualifications amidst concerns about the decoupling of AS and A levels. Current reforms have also led to changes in models of assessment including those involving inclusion and balance of examined and non-examined assessment. Gill summarises the processes undertaken by Oxford, Cambridge and RSA (OCR) to moderate coursework and controlled assessment. He discusses the aims and processes of moderation as well as the principle of fairness for all candidates.

Assessment strategies are important for the development of the reformed qualifications and validity is central to any assessment strategy. Shaw and Crisp make a timely contribution by reflecting on Cambridge Assessment's validation research which led to the development of a framework for evidencing validity in large-scale, high-stakes examinations. Over the last five years the framework has been amended and strengthened.

The final article looks to potential future developments in a different field. Zanini and Dhawan focus on statistical and Computer Science techniques which have been developed to analyse text data. They discuss new sources of text data such as text messaging, social media activity, blogs and web searches in the context of the Big Data trend. This is an expanding field and offers the opportunity for new areas of research and new methodologies.

Sylvia Green *Director of Research*

A level History: Which factors motivate teachers' unit and topic choices?

Simon Child, Ellie Darlington and **Tim Gill** Research Division

Introduction

During periods of curriculum and qualifications reform, debates typically centre on establishing the fundamental content, skills and competencies that students should possess in different subject domains. Currently, England is undergoing a period of reform in secondary education, where significant changes to course content are to be introduced, alongside structural changes to general qualifications (Department for Education [DfE], 2010).

At A level, the changes to subject content have been guided by Higher Education. *The Smith Report* (2013), commissioned by The Office of Qualifications and Examinations Regulation (Ofqual), made recommendations with regards to content changes for the reformed A level qualifications in 15 subjects. For the subject of History, the Smith Report recommended that A levels should cover at least a 200 year period, and should focus on more than one state. Perhaps in response to difficulties in defining appropriate historical content for the successful transition to university (Hibbert, 2006), there was little direction in terms of specific content areas. Schools have historically been offered flexibility in the topics they cover at A level History. For example, in the current Oxford, Cambridge and RSA (OCR) A level (Specification A) there are 16 possible unit combinations available to students, and a range of different topics within each unit can be taught. Other exam boards offer fewer options in terms of unit choice, but a greater range of topic options within units.

Although at first glance this course flexibility would appear to encourage the teaching of a wide range of historical topics, recent research has suggested that schools tend to teach narrow historical periods. For example, in their analysis of the unit and topic choices taken within one History A level, Child, Darlington and Gill (2014) found that schools were more likely to choose units that focused on modern History and within these units, centred on specific twentieth century topics.

The flexibility inherent in A level History qualifications means that teachers have to negotiate competing factors that may influence topic, unit or qualification choices. First, the study of History can serve several purposes for students (see Barton & Levstik, 2004, for a review). For example, Harris (2013) argued that History operates for communities in much the same ways as memory does for individuals, in that it facilitates more informed decision making. Secondly, students are also likely to be engaged by different topics, as they may identify with different geographical regions or cultural backgrounds. As changes to qualifications are introduced, teachers may be aware that the introduction of new topic areas may be problematic if students do not identify with them (Elwood, 2012). Thirdly, students are likely to be influenced by the school they attend (Nelson, Morris, Rickinson, Blenkinsop & Spielhofer, 2001; Vidal Rodeiro, 2007). For example, Vidal Rodeiro (2007) found that independent school students were more likely to choose 'traditional' subjects (including History) compared to comprehensive school students. Fourthly, teachers may have their own areas of interest or expertise which may influence the topics they teach. This expertise may be developed through experience teaching the topic in school, by previous degree level study, or by personal interest (Chandra, 1987). However, even in cases where teacher expertise is not as well developed, the availability of high-quality resources may encourage teachers to select certain historical topics. In times of curriculum change, teachers have to re-assess these factors for the benefit of the school and the student.

Aims of the current study

Given the tensions outlined, it is surprising that little previous research has examined which factors influence unit and topic decision choices in History. The present study aimed to use questionnaire data derived from heads of History departments to analyse the motivations underpinning the unit and topic choices for an A level History course. A second aim was to analyse whether the Heads of Department from different school types had different influences underlying their choices.

Method

The data for the present study were collected as part of a larger research study that aimed to investigate the scope of historical topics taught at A level (see Child et al., 2014). This research involved the statistical analysis of question-level data for an A level History course, and a questionnaire sent to heads of History departments in schools. An overview of the questionnaire method is presented below.

Participants

Centres that took OCR A level History in June 2013 were contacted by telephone, and asked to provide the full name and contact details for the head of the History department or equivalent. The Heads of Department were then emailed and invited to fill out the questionnaire, which they could access via a weblink. As an acknowledgement of their time, they were offered the opportunity to enter into a prize draw.

Of the 638 Heads of Department contacted, 90 returned the questionnaire (a return rate of 14%). Participants had a mean of 6.71 years of experience (*SD* = 6.21 years) as Head of Department at the centre where they were currently employed. The centres had spent a mean of 11.89 years teaching OCR A level History (*SD* = 6.25 years).

Eighty-five of the participants provided information about the type of school where they were teaching. Fifty-two of the centres were state schools, and 33 were independent schools. The percentage of schools in this sample that were independent (39%) is slightly higher than the

overall percentage of independent schools that take OCR History (34%). However, we deemed that this sample was broadly representative of the total population of centres that offered OCR A level History in 2013.

Questionnaire development

The questionnaire was developed by members of the research team in collaboration with the OCR General Qualifications Reform Subject Team. The questionnaire comprised three sections related to the decision process underlying unit and topic choices for A level History. The first section asked participants for details of their centre and teaching experience. The second section asked them about their role in the decision process of making unit and topic choices (e.g., if it was their decision alone or decided after discussion with colleagues). The third section asked them to rate how important 11 factors are when deciding the unit and topic choices for A level History.

Piloting

Before the questionnaire was distributed to the participants, a draft version was checked by the OCR Subject Team for History, to ensure that appropriate terminology and question response choices were included. The questionnaire was then sent to a pilot participant, who was a Head of Department for History. The pilot participant was asked to check the questionnaire for anything that they felt would not be understood by participants, and errors in spelling or grammar. They were also asked if there were responses that could be added to any of the questions. Once the recommended changes were made, a weblink for the final version was sent to the main cohort.

Results

The results are presented in two sections. The first section relates to the unit choice decisions, while the second section relates to the topic choices within each unit.

Factors affecting unit decisions

Overall, 67.8% of the Heads of Department reported that the decision on the unit choices that would be offered to students was made after discussion with other teachers in their department, while 21.1% of Heads of Department reported that they alone made unit choice decisions. This pattern was similar across the two school types. Overall, it was rare for decisions on unit choice to be made on a class-by-class basis (4.4%). The 6.7% of Heads of Department who selected 'other' explained that they made unit decisions after some form of student consultation. Again, this strategy for unit selection was distributed evenly between the school types.

Table 1 shows how important 11 factors were in deciding the unit choice decisions. Overall, the two most important factors that determined schools' unit topic choice were teacher expertise and student engagement, with over 81% of Heads of Department deeming these factors as important. Other important factors included the availability of paper-based resources and breadth of topics studied across the course. Interestingly, multimedia-based resources were regarded as less important by Heads of Department, as only 10% of them reported them to be important, and 61.1% regarded them as not at all important. The importance of having effective teaching resources for teachers was

Table 1: Unit decision factors overall and by school type

Factor	*Percentage of participants*											
	Important			*Somewhat important*			*Not at all important*			*Don't know*		
	Overall	*State*	*Independent*	*Overall*	*State*	*Independent*	*Overall*	*State*	*Independent*	*Overall*	*State*	*Independent*
Expertise of the A level teachers within the History Department	82.2	86.8	72.7	15.6	11.3	24.2	1.1	0.0	3.0	0.0	0.0	0.0
Paper-based resources available in the History Department	45.6	56.6	30.3	40.0	32.1	48.5	13.3	9.4	21.2	0.0	0.0	0.0
Multimedia resources available in the History Department	10.0	15.1	3.0	27.8	35.8	15.2	61.1	47.2	81.8	0.0	0.0	0.0
Resource availability	23.3	32.1	12.1	44.4	47.2	42.4	31.1	18.9	45.5	0.0	0.0	0.0
Resource quality	26.7	34.0	18.2	41.1	43.4	36.4	28.9	18.9	42.4	0.0	0.0	0.0
Breadth of topics studied across the course	41.1	47.2	36.4	50.0	47.2	48.5	6.7	3.8	12.1	1.1	0.0	3.0
Link between time periods studied across the course	34.4	35.8	36.4	45.6	45.3	42.4	18.9	17.0	21.2	0.0	0.0	0.0
Link to the previous educational level	11.1	15.1	6.1	40.0	49.1	27.3	44.4	32.1	60.6	2.2	1.9	3.0
Student engagement with course content	81.1	90.6	66.7	16.7	7.5	30.3	0.0	0.0	0.0	0.0	0.0	0.0
Perceived ease of unit content	14.4	20.8	6.1	52.2	54.7	48.5	31.1	20.8	45.5	0.0	0.0	0.0
Links to A level History courses previously taught at the school	6.7	9.4	0.0	16.7	22.6	3.0	67.8	58.5	87.9	4.4	5.7	3.0

supported by the finding that resource availability and resource quality was rated as important by approximately a quarter of respondents. The factor that was rated overall as least important was links to A level courses previously taught at the school, with over two thirds of participants rating it as not at all important.

There were differences found between school type with respect to which factors were most important in making unit decisions. Whilst 90.6% of state school Heads of Department deemed student engagement to be important, only 66.7% of independent school Heads of Department thought this was an important factor. To test whether this difference between school type was statistically significant, Fisher's Exact test was run with the categorical variables of *school type* (state versus independent) and *important* or *other*[1]. Fisher's Exact test was found to be significant (p = .018) suggesting that student engagement was significantly more important in state schools, when deciding which units to select. Paper-based resources were also more important for state schools compared to independent schools, with a difference between them of 26.3 percentage points (Fisher's Exact, p = .034). Similarly, Heads of Department at state schools perceived resources as being more important in making unit decisions compared to independent schools. At state schools, 32.1% and 34.0% of Heads of Department regarded resource availability and resource quality respectively to be important compared to 12.1% and 18.2% at independent schools. However, the difference between school type for resource availability was only approaching significance (Fisher's Exact, p = .073), and the difference for resource quality was non-significant (Fisher's Exact, p = .265). State school Heads of Department also rated ease of course content to be important more than independent school Heads of Department (20.8% versus 6.1%), although again this difference was only marginally significant (Fisher's Exact, p = .083). Expertise of the teachers within the department was thought to be important overall, but there was no significant difference found between school types (Fisher's Exact, p = .283).

1. The other category comprised of centres that had rated student engagement as *Somewhat important*, *Not at all important*, or *Don't know*. T.C. Benton (personal communication, 7 May 2014).

Factors affecting topic decisions

Overall, 68.9% of the Heads of Department reported that the decision on the topic choices that would be offered to students was decided after discussion with other teachers in their department, while 23.3% of Heads of Department reported that they alone made unit choice decisions. This pattern was again similar across the two school types. It was rare for decisions to be made on a class-by-class basis (7.8%) and this strategy was distributed evenly between state and independent schools.

Table 2 shows how important 11 factors were in deciding the topic choice decisions. Overall, expertise of the A level teachers and student engagement with course content were regarded as the two most important factors in making topic decisions, with over 71% of Heads of Department reporting these factors to be important. Other factors that were highly rated as important by Heads of Department included paper-based resources and breadth of topics studied across the course. Relative to paper-based resources, multimedia-based resources were regarded as less important by Heads of Department, with only 13.3% rating them as important and over half rating them as not at all important. Other

Table 2: Topic decision factors overall and by school type

Factor	*Percentage of participants*											
	Important			*Somewhat important*			*Not at all important*			*Don't know*		
	Overall	*State*	*Independent*	*Overall*	*State*	*Independent*	*Overall*	*State*	*Independent*	*Overall*	*State*	*Independent*
Expertise of the A level teachers within the History Department	78.9	86.8	63.6	13.3	5.7	27.3	2.2	1.9	3.0	0.0	0.0	0.0
Paper-based resources available in the History Department	42.2	54.7	24.2	38.9	30.2	48.5	13.3	9.4	21.2	0.0	0.0	0.0
Multimedia resources available in the History Department	13.3	20.8	3.0	27.8	34.0	15.2	52.2	39.6	72.7	0.0	0.0	0.0
Resource availability	22.2	32.1	9.1	43.3	45.3	39.4	28.9	18.9	42.4	0.0	0.0	0.0
Resource quality	23.3	34.0	9.1	41.1	41.5	39.4	30.0	18.9	45.5	0.0	0.0	0.0
Breadth of topics studied across the course	35.6	37.7	36.4	51.1	54.7	42.4	5.6	1.9	9.1	1.1	0.0	3.0
Link between time periods studied across the course	22.2	24.5	21.2	52.2	52.8	48.5	17.8	15.1	21.2	0.0	0.0	0.0
Link to the previous educational level	10.0	13.2	6.1	37.8	47.2	24.2	45.6	34.0	60.6	2.2	1.9	3.0
Student engagement with course content	72.2	83.0	54.5	22.2	11.3	39.4	0.0	0.0	0.0	0.0	0.0	0.0
Perceived ease of unit content	11.1	18.9	0.0	52.2	56.6	45.5	30.0	20.8	42.4	0.0	0.0	0.0
Links to A level History courses previously taught at the school	4.4	7.5	0.0	20.0	30.2	1.0	62.2	47.2	84.8	4.4	5.7	3.0

factors of less importance to Heads of Department included links to previous A levels taught within the department and links to the previous educational level, both of which were rated as not at all important by over 45% of Heads of Department. The importance of having effective teaching resources for teachers was again supported by the finding that resource availability and resource quality was rated as important by over a fifth of Heads of Department.

This pattern of results by school type was similar to the findings of the unit level decision factors. Although student engagement was reported to be important overall, significantly fewer independent school Heads of Department perceived this factor to be important compared to state Heads of Department (Fisher's Exact, $p = .008$). Interestingly, while at the unit level there was no significant difference found in the perception of the importance of teacher expertise between school types, at the topic level, a significant difference was found (Fisher's Exact, $p = .020$), with Heads of Department at state schools more likely to rate teacher expertise as important compared to independent school Heads of Department. For the factors that related to resources (paper-based resources, multimedia-based resources, resource availability, and resource quality), state school Heads of Department were more likely to rate these as important compared to independent school Heads of Department; for all of these factors there was a difference of 17.8 percentage points or greater. Fisher's Exact test revealed that the differences for all the resource-related factors were significant (paper-based resources, $p = .018$; multimedia-based resources, $p = .038$; resource availability, $p = .039$; resource quality, $p = .032$).

Heads of Department of independent schools perceived factors that were linked to the students' or schools' past experience with qualifications (link to previous level of education, and links to previously taught A level History courses) as not important more often (60.6% and 42.4% respectively) than state Heads of Department (34% and 20.8%). In both cases, Fisher's Exact was significant (link to previous education, $p = .033$; links to previously taught A level History courses, $p = .001$).

Finally, state school Heads of Department also rated ease of course content to be important more than independent school Heads of Department (18.9% versus 0%; Fisher's Exact, $p = .14$).

Discussion

The present study aimed to explore the importance of 11 factors that History departments might consider in the selection of A level History units and topics. This was in the context of previous research that had found a tendency for schools to select units and topics that covered similar historical periods and geographical locations. The questionnaire data revealed that some factors were more important than others and that in some cases their importance was influenced by school type.

The factor that was rated as most important overall at the unit and topic levels was teacher expertise. However, while there was no statistical difference observed between state and independent schools for this factor at the *unit* level, there was a difference observed at the *topic* level. A similar finding was observed at the topic level in terms of the importance of resources, with state school Heads of Department rating resources as more important compared to independent schools. It is likely that these factors are related. State school Heads of Department perhaps need more assurance that teachers are comfortable with the topic that they have been asked to teach. This consideration may be in response to the size of the state school departments relative to independent schools (DfE, 2011). The availability and quality of resources, however, may be a mediating factor for teachers that have less experience with particular topics, or in times of curriculum change (Child, Devine & Wilson, 2013). Interestingly, paper-based resources appeared to be more highly valued by the Heads of Department, which contrasts to the increasingly multimedia driven delivery in other subject areas (Bauer, 2005; Hooper & Rieber, 1995). One interpretation is that the focus of History on the analysis of the relation between evidence and the construction of historical accounts (Barton & Levstik, 2003) lends itself to more kinaesthetic, physical representations of sources.

Student engagement was rated overall as the second most important factor in making unit and topic choices. However, it is unclear as to why independent schools were less likely to rate student engagement as important compared to state schools. One potential reason may be that state schools are typically confronted with a more varied student population in terms of cultural background (DfE, 2014). Whilst students play an active role in the construction of their own knowledge and relate this knowledge to their experience, they also inhabit the pedagogical framework constructed by teachers. As part of this 'social' or 'didactic' contract (Brousseau, 1984; Schubauer-Leoni, Bell, Grossen & Perret-Clermont, 1989) students rely on the teacher to make decisions related to course content and delivery. For state school teachers, this concern may be at the forefront of their thinking when deciding which topics to teach. However, in the present study, a similar number of state and independent school Heads of Department reported to consulting students before making unit decisions in the present study. Future research is required to determine the process of student consultation for courses where unit choices are available. For example, it may be the case that while some teachers may consult students to merely confirm unit choices, other teachers may be more open to student-level decision making at an early stage. An analysis of these processes may reveal differences in teacher approaches to the initial building of course content in collaboration with students.

Breadth of topic coverage was identified as important for the majority of Heads of Department. This contrasts with the Child et al. (2014) finding that at the unit level, schools are more likely to teach topics that cover similar time periods and subject matter. Teachers may be looking for internal coherence within the qualification, so that maximum depth can be achieved in topic areas that are of interest to students (or that they can identify with, Harris, 2013). This qualification-level coherence could also be in response to teaching time pressures, or the assumption that for students who intend to study History at university, content knowledge is less important than skill development (Smith, 2013). Indeed, in the first year at university, courses focus on key skills which are then applied to historical periods. For example, at the University of Exeter, the three compulsory modules taken by first year undergraduates relate to the development of skills in referencing appropriately, thematic analysis of sources, working independently, and understanding recurring themes in History (University of Exeter, 2014). These core modules are supplemented by modules on particular historical periods and topics.

The desire for within–qualification content coherence observed in the present study does not appear to be matched by the intention to match up the study of topics between GCSE and A level. It appears then that for History, particular historical content knowledge is not a prerequisite for effective transfer to the next educational stage. This is interesting as it contrasts with the new National Curriculum's emphasis on

'chronologically secure knowledge' and recent political rhetoric on "Our island [UK's] story" (Gove, 2010).

The recommendations of the *Smith Report* (2013) outline a qualifications framework for A level History that allows students to cover a sufficient breath of historical eras, but with few limitations on specific topics. In some cases, this 'enforced optionality' approach to A level History qualifications will mean a period of adjustment (Child et al., 2014), with new topics introduced for the first time to meet the demands of the qualification. For example, the newly accredited OCR A level History course (OCR, 2014) comprises four compulsory units based on geographical factors (British and non-British History) and skills development (thematic understanding and a topic-based essay). Within both the British and non-British History units, there are over 21 topics that can be studied, with newly introduced areas of study including *The Rise of Islam (c. 550–750)*, *Japan (1853–1937)*, and *Charlemagne (768–814)*. Future research could explore how students before and after the reforms perceive A level qualifications in terms of their aims and their usefulness for undergraduate study. It would also be interesting to explore the implicit assumption that the skills developed during the study of A level History are largely in isolation to the context provided by the historical period studied. Analysing students' perceptions of the skills they learned studying History may reveal that the topic areas they identified with most were more effective in developing their analytical and written abilities.

Acknowledgements

We wish to thank our Research Division colleagues, Tom Benton, Sylvia Green, Tom Bramley, Lucy Chambers and Irenka Suto, and Mike Goddard from OCR, for their helpful advice on this article. We also wish to thank Jo Ireland, Research Division, for her administrative assistance during the study. Finally, we are grateful to the participants for engaging with this research.

References

Barton, K.C. & Levstik, L.S. (2003). Why don't more history teachers engage students in interpretation? *Social Education*, *67*(6), 358–361.

Barton, K.C. & Levstik, L.S. (2004). *Teaching History for the Common Good*. Mahwah, NJ: Lawrence Erlbaum.

Bauer, J. & Kenton, J. (2005). Toward technology integration in the schools: Why it isn't happening. *Journal of Technology and Teacher Education*, *13*(4), 519–546.

Brousseau, G. (1984) The crucial role of the didactic contract in the analysis and construction of situations in teaching & learning mathematics. In J. G. Steiner (Eds.), *Theory of Mathematics Education* (pp.110–119). Bielefeld, Germany: University of Belefield.

Chandra, P. (1987). How do teachers view their teaching and the use of teaching resources? *British Journal of Educational Technology*, *18*(2), 102–111.

Child, S.F.J., Darlington, E., & Gill, T. (2014). An analysis of the unit and topic choices made in an OCR A level History course. *Research Matters: A Cambridge Assessment Publication*, *18*, 2–9.

Child, S.F.J., Devine, A., & Wilson, F. (2013). *"It's gold dust." Teachers' views on curriculum support resources*. Cambridge: Cambridge Assessment.

DfE. (2010). The Importance of Teaching – the Schools White Paper. Retrieved from https://www.education.gov.uk/publications/standard/publicationDetail/Page1/CM%207980

DfE. (2011). Schools, pupils, and their characteristics. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/219064/main_20text_20sfr122011.pdf

DfE. (2014). Schools, pupils, and their characteristics. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/335176/2014_SPC_SFR_Text_v101.pdf

Department for Education and Skills. (2007). *Diversity and citizenship curriculum review*. London: DfES.

Elwood, J. (2012). Qualifications, examinations and assessment. Views and perspectives of students in the 14–19 phase on policy and practice. *Cambridge Journal of Education*, *42*(4), 497–512.

Gove, M. (2010). Speech to the Conservative Party Conference in Birmingham. Retrieved from http://centrallobby.politicshome.com/latestnews/article-detail/newsarticle/speech-in-full-michael-gove/

Harris, R. (2013). The place of diversity within history and the challenge of policy and curriculum. *Oxford Review of Education*, *39*(3), 400–419.

Hibbert, B. (2006). *The articulation of the study of history at General Certificate of Education Advanced Level with the study of history for an honours degree*. Unpublished doctoral thesis, University of Leeds.

Hooper, S., & Rieber, L. P. (1995). Teaching with technology. In A. C. Ornstein (Eds.), *Teaching: Theory into practice*, (pp.154–170). Needham Heights, MA: Allyn & Bacon.

Nelson J., Morris M., Rickinson M., Blenkinsop S. & Spielhofer T. (2001). *Disapplying National Curriculum subjects to facilitate work-related learning at Key Stage 4: An evaluation* (DfES Research Report 293). London: DfES.

OCR. (2014). *OCR Level 3 Advanced GCE in History A (H505) Specification*. Retrieved from http://www.ocr.org.uk/Images/170128-specification-accredited-a-level-gce-history-a-h505.pdf

Schubauer-Leoni, M.L., Bell, N., Grossen, M., & Perret-Clermont, A.N. (1989). Problems in assessment of learning: The social construction of questions and answers in the scholastic context. *International Journal of Educational Research*, *13*(6), 671–684.

Smith, M. E. (2013). *Independent Chair's report on the review of current GCE 'specification content' within subject criteria: A report to Ofqual*. Retrieved from http://ofqual.gov.uk/qualifications-and-assessments/qualification-reform/a-level-reform/

University of Exeter. (2014). Undergraduate study: B.A. History Retrieved from http://www.exeter.ac.uk/undergraduate/degrees/history/historyexe/#Programme-structure.

Vidal Rodeiro, C.L. (2007). *A level subject choice in England: Patterns of uptake and factors affecting student preferences*. Cambridge: Cambridge Assessment.

Context led Science courses: A review

Frances Wilson Research Division, **Steve Evans** OCR and **Sarah Old** OCR

Introduction

Internationally, there is growing concern about secondary Science education. In many developed countries[1], uptake of Science subjects has been falling (Bennett, Gräsel, Parchmann, & Waddington, 2005), leading to fears that there will be a shortage of people with the scientific skills and knowledge needed in the twenty-first century. The lack of uptake has been attributed to various causes. In particular, Science curricula are often considered to suffer from an overload of content, leading to the perception that Science subjects are among the most difficult. Furthermore, students have difficulties connecting the isolated facts which they are taught, and do not develop coherent mental schema. Content is often presented in an abstract manner that is remote from students' everyday experiences, so that many students do not understand why they should learn the materials which they are studying, and frequently fail to do so. As a result, students have difficulty applying scientific concepts in a context beyond the one in which they were taught that concept (Bennett et al., 2005; Fey, Gräsel, Puhl, & Parchmann, 2004; Gilbert, 2006; Nawrath, 2010; Pilot & Bulte, 2006). In this article we examine one approach to Science education: context led Science courses, which have been developed as a result of these concerns.

Context led Science

Traditional Science courses can be said to be "concept led", that is, they are structured from the perspective of a scientist, with scientific concepts organised in a way that makes sense to a scientist (Reiss, 2008). In contrast, a context led approach can be characterised by the "use of contexts and applications of Science as the starting point for developing scientific understanding." (Bennett et al., 2005, p.1523). A structure based in contexts may relate better to students' own knowledge about the world. For example, a concept led Biology course might structure the course into topics such as 'Biochemistry' or 'Cell Biology', whereas a context led course might use the context of crop production and global food security to introduce plant transport, reproduction and the biochemistry of photosynthesis. From this starting point, a context led course would then revisit other biochemical topics in other contexts at a later point in the course.

Context led Science courses aim to address the problems associated with traditional Science courses by breaking down boundaries between school Science and everyday contexts to increase the social and cultural relevance of Science for students, by making the relationship between social issues and scientific knowledge more prominent. It is hoped that this sort of approach will provide greater access to Science education for groups of learners who traditionally do not participate fully in post-compulsory Science education, such as certain cultural minorities, or girls. Furthermore, it has been argued that by relating school Science to authentic scientific contexts, students may develop a greater understanding of the range of scientific careers which are available, potentially increasing uptake of Science subjects (Lubben & Bennett, 2008).

What is a context?

Although the term "context led" is commonly used, "context" may have several meanings. For example, at its widest, it might refer to the social and cultural environment in which the student, teacher and institution are situated, or, more narrowly, refer to the application of a scientific theory (Bennett et al., 2005). Giamellaro (2014) proposes that the process of contextualising knowledge involves forming specific connections between the content knowledge which is taught, and an authentic environment in which the content can be relevantly applied or illustrated. Decontextualised knowledge, on the other hand is typically only used in scholastic environments, and is abstracted away from the content knowledge as it is typically used in practice (p.2849). For example, knowledge of intermolecular bonds, such as hydrogen bonds is decontextualised, but can be contextualised when linked to polymer properties. A context led course uses the authentic environment as a starting point for teaching and learning, whereas a traditional concept led course starts with decontextualised knowledge, which might (but not necessarily) then be applied to a context. Giamellaro further distinguishes between learning *with* context, using a second hand context, such as a case study, and learning *in* context, such as an internship. However, what is considered to be an authentic environment may vary. For example, it is not clear whether a hypothetical case study, such as a boy who has had a stroke, can be considered to be truly authentic, if the case study has been designed specifically for educational purposes.

Attributes of contexts

Gilbert (2006), following Duranti and Goodwin (1992), proposes that an educational context can have four attributes. For example, a context used to study the Chemistry of global warming (focal event) would have the features shown in Table 1.

The contexts used in Science courses may include social, economic, environmental, technological and industrial applications of Science. Some courses select contexts which are directly relevant to students' personal circumstances, while others may focus on societal/community issues, or contexts which are relevant from a vocational perspective. (Kazeni & Onwu, 2013). In general, for younger students, contexts which have direct applications to students' lives are typically used, whereas for older and more advanced students, contexts which explore 'what scientists do' may be more common. Pilot and Bulte (2006) and Gilbert (2006), in the

1. In England, there has been a recent increase in the percentage of A level entries for Science (http://sciencecampaign.org.uk/?p=12878), although since this follows a period of decline over several decades (Bennett, Lubben, & Hampden-Thompson, 2013), this growth needs to be sustained over several years to allow uptake to recover fully.

Table 1: Attributes of an educational context (adapted from Gilbert, 2006, p.961)

Attribute	*Example*
Setting: Where, when, how is the focal event situated?	The setting is the specific example of the focal event. The focal event is the general phenomenon of global warming, manifest throughout the world in different settings,
Behavioural Environment: What do people do in this situation; what actions do they take?	People take various measures to reduce the production of relevant gases, and remove those already in the atmosphere.
Language: What language do people use to speak about their actions?	The molecular structures of relevant gases are discussed, with a particular emphasis on the way that internal vibrations lead to the effects that are observed.
Background Knowledge: What is the background knowledge of those who act?	The need for a general education about molecular structure and energy conversion is required.

context of Chemistry education, outline four criteria needed for the successful use of context:

1. Students must value the setting, and recognise that it falls within the domain of Chemistry. It must arise from the everyday lives of the students, or social issues and industrial situations that are of contemporary importance to society.
2. The behavioural environment must include problems that are clear exemplifications of chemically important concepts, so that students engage in activities from the domain of Chemistry, such as experimental laboratory skills.
3. Learners should be enabled to develop a coherent use of specific chemical language which is brought into focus by the behavioural environment.
4. The behavioural environment and the language used to talk about it should relate to relevant extra-situational, background knowledge, building productively on that prior knowledge.

(Gilbert, 2006, p.961)

Additionally, across a course as a whole, curriculum developers must plan contexts which allow students to revisit scientific concepts, albeit from a different perspective, in a way which allows students to build up their understanding of scientific topics. These contexts should enable students to make analogies between contexts, so that it is clear that concepts can be transferred to these new contexts (Gilbert, 2006).

However, for curriculum developers, there may be significant challenges in meeting these criteria, particularly if a Science course is to be taught to students from diverse backgrounds. Not all contexts are suitable for use in every context. For example, Kazeni and Onwu (2013) give the following context which was used successfully in a context led course on genetics in South Africa:

> *Mr. and Mrs. Sizwe have been married for twelve years. They have four daughters, and no son. According to Mr. Sizwe's custom, not to have a son means that there would be no heir to succeed him. Mr. Sizwe decided to consult his elders about his situation. After consulting with them, he decided to take on a second wife who would bear him a son. To his dismay, the second wife gave birth to a girl.*
>
> *The question is: How can the situation about sex determination be resolved scientifically?*
>
> (Kazeni & Onwu, 2013, p.55)

Although students in South Africa may be able to relate to this context, it seems unlikely that it could be successfully used in secondary schools in England, where students may not be familiar with the cultural need for a son, nor the practice of taking a second wife. Furthermore, Taasoobshirazi and Carr (2008) note that if students become too emotionally engaged with the context, then this might distract them from learning the relevant scientific concepts. Similarly, contexts which are too complicated, or provide too much interesting, but not relevant, information might be confusing. However, if such contexts are not part of the everyday lives of students, or they are not engaged with particular social issues, then they may not be engaging enough.

Pedagogy in context led Science courses

Context led Science courses are strongly associated with particular pedagogical approaches to teaching Science. Bennett et al. (2005) note that context led courses typically use a "spiral" curriculum, in which students encounter the same concepts, albeit from a different perspective across multiple contexts at different stages of the course. This may help students to connect otherwise isolated facts, and develop coherent mental schema. Revisiting the same concept in different contexts allows students the opportunity to transfer their application of a concept to different contexts. For example, in the Salters A level Chemistry course, chemical equilibrium is introduced initially in the unit "The Atmosphere", in relation to reversible reactions to explain the role of carbon dioxide in the oceans. It is later developed further in "The Steel Story", by looking at redox reactions, and then revisited in "Aspects of Agriculture" to explain ion-exchange equilibria. Towards the end of the course, the concept is extended to more complex situations, such as pH and buffer solutions.

Context led Science courses are generally characterised by the adoption of a student-centred approach to teaching, which requires students to engage in meaningful activities, rather than rote learning (Overman, Vermunt, Meijer, Bulte, & Brekelmans, 2012). For example, King and Ritchie (2013) describe a project undertaken by 11th grade Chemistry students in Australia, in which students investigated the water quality in their local creek by carrying out tests on water from the creek taken from three locations. Students were required to conduct background research on each test (e.g., for Dissolved Oxygen, pH, turbidity, Biochemical Oxygen Demand, salinity, and faecal coliforms), and then report on the overall water quality based on their understanding of the various tests, and the chemical concepts underpinning these tests. The use of a local context helped to make the project meaningful to students, while their research into each water quality test helped them to develop the appropriate language to talk about the underlying chemical concepts.

Such student-centred approaches to the organisation of the curriculum have their roots in constructivist theories which emphasise the importance of learners actively constructing their knowledge. In particular, constructivist approaches are based on the principle that students must be actively involved if they are to achieve understanding, and that students and their ideas should be respected, so that teaching allows students to use what they already know, and can address difficulties that result from a naïve understanding of scientific ideas (Gilbert, 2006; Gilbert, Bulte, & Pilot, 2011). Educational constructivists would further argue that a traditional, transmission-based approach[2] to

2. Transmission-based approaches are centred around the transmission of information from teachers to students, in which the teacher structures and organises the information for the students. As such, it is a teacher-centred approach (Overman et al., 2014).

teaching is unlikely to lead to students developing a meaningful understanding of the content, and that this is only achieved by teaching methods which allow students to engage with the material (Overman et al., 2012). A successfully applied context allows students to use their own background knowledge and understanding of the context, helping them to make the scientific concepts associated with the context meaningful. However, one consequence of this link between pedagogy and context led teaching is that teachers who switch from a traditional concept led course may also need to learn new pedagogical approaches at the same time. As a result, this may make a shift towards context led courses particularly demanding for teachers. Furthermore, this type of pedagogical approach may be challenging when classes are large and diverse in terms of prior knowledge and experience, and future goals (Gilbert, 2006).

Despite the relationship between constructivist pedagogy and context led Science courses, Peşman and Özdemir (2012) note that it is possible to use student-centred learning approaches in a traditional concept led course, and transmission based teaching in a context led course. Indeed, in a short term (five weeks of teaching time) study, they found that a student-centred, active learning approach was more effective for a concept led course, compared to a context led course. Somewhat surprisingly, transmission methods of instruction seemed to be more effective with context led courses. It is possible that introducing only one innovation (a change in teaching method or a move to a context led course) was most effective over this time period, because students and teachers were able to adapt to one innovation but not both. Furthermore, even if teaching activities (such as inquiry based projects or student discussions) which are promoted by constructivist approaches are not used, the use of a context to introduce a topic may help students to engage with the topic, and understand why the scientific concept is relevant to their everyday lives. Similarly, a concept led course which is taught in a way that recognises students' prior knowledge and experiences may be more successful than one which uses a traditional transmission pedagogy.

Models of embedding context

Given the range of different understandings of the term 'context', combined with potentially different approaches to teaching context led courses, context led Science courses should not be considered a homogenous group. Gilbert (2006) and Gilbert et al. (2011) propose four models for context led courses. Each model represents a different way in which context is embedded in the course.

Model 1: Context as the direct application of concepts

This model represents what is typical of many concept led courses. Concepts are decontextualised, and typically presented as abstractions. Contexts are only subsequently introduced, typically allocated little time, and not used in assessment. Such courses are not generally considered to be context led, because contexts are introduced after concepts.

Model 2: Context as reciprocity between concepts and applications

Contexts are selected as a means through which concepts can be taught, and juxtaposed with the relevant concepts. This model may be considered to be more context led than Model 1, because it does provide a setting and behavioural environment which students may use as a framework, and may enable them to relate what is being learned to their prior knowledge. However, the lack of a clear rationale for the integration of contexts may mean that students do not relate strongly to the context. Furthermore under this model of contextualisation, the degree to which concepts are repeatedly recontextualised may vary.

Model 3: Context as provided by personal mental activity

This model focuses on learners who are working as individuals, typically from a book or online courses. It is characterised by the use of a narrative to frame historical events, which may allow students to empathise with the participants in the narrative. For example, students may study the events leading to an important scientific discovery. This model lacks a social dimension.

Model 4: Context as the social circumstances

In this model students and teachers work together on an enquiry into a topic which is considered of importance to the lives of their community. Learning takes place as students experience a setting, and by participating in interactions with members of their community.

Examples of context led courses

In this section we describe four different context led Science courses, which were developed with similar aims, but in different educational contexts.

Salters – England

The Salters project began in the early 1980s, when a group of teachers and Science educators met at the University of York to discuss how Chemistry education could be made more appealing to secondary school students. Since then, the Salters project has expanded to include Biology and Physics as well as Chemistry, leading to the development of Science courses for students aged 11–18 (Bennett & Lubben, 2006), and has been used as a model for context led courses internationally (Parchmann et al., 2006). All Salters courses are based on the same design criteria, namely that the contexts and concepts selected for study should enhance students' appreciation of how Science contributes to their lives, or the lives of others around the world, and to help them understand the natural environment better (Bennett & Lubben, 2006). The courses use a spiral curriculum, such that scientific concepts are re-visited in different contexts throughout the course.

Here we focus on the Salters A level courses, which were first developed for Chemistry in the late 1980s (Bennett & Lubben, 2006), for Physics (Salters-Horners) in the early 1990s (Institute of Physics, 2003), and in the early 2000s for Biology (Salters-Nuffield) (Reiss, 2005). The Biology course was introduced later than the Physics or Chemistry course because there are in general fewer concerns about the uptake of Biology, although Biology teaching has been criticised for using activities which require little student involvement, and do not include enough practical work (Reiss, 2005). All three A level courses have been developed as a partnership between the University of York and exam boards (Oxford, Cambridge and RSA (OCR) for Chemistry, Pearson-Edexcel for Biology and Physics). Pilot and Bulte (2006) argue that the integration of the Salters courses with national examinations facilitated uptake of the courses, and was critical to their success. The courses are distinct from traditional concept led A level courses due to the use of a spiral curriculum, based around different contexts, and the use of personal investigations conducted by students. At AS level, these include a report based on a literature review or a visit to a site (e.g., zoo, local chemical industry) (Astin, Fisher, & Taylor, 2002; Dunkerton, 2007) and at A2, an extended experimental investigation (Lewis & Scott, 2006).

Until 2008, separate specifications and assessments were developed for the Salters context led courses, and traditional concept led courses. However, when A level specifications were re-developed for first teaching from 2008, the Biology and Physics courses (both Pearson-Edexcel) were designed so that both the Salters and the traditional courses shared the same assessment, but teachers could choose whether to teach the content using a context or concept led approach[3]. It is not clear whether this approach to assessment is successful: sample assessment materials for these courses (Edexcel, 2014) seem to predominantly assess students using concept led questions, with the exception of the questions based on a scientific article. The use of concept led questions may not allow students who have followed a context led course to fully demonstrate the skills which they have acquired, and the content and form of the assessment is likely to influence teaching and learning (Pilot & Bulte, 2006). However, Braund, Bennett, Hampden-Thompson, and Main (2013) found no significant difference in the marks obtained by students following a concept led course compared to a context led course, suggesting that neither teaching approach disadvantages students. In this study, centres were classified according to a combination of self-report, access to context led teaching resources and historical teaching approach, so it is likely that there was some diversity of approach within both the concept and context led groups, which possibly reduced any difference in outcomes between the two groups.

Bennett et al. (2005) investigated A level Chemistry teachers' views on the OCR Salters A level Chemistry course and the traditional concept led Chemistry course. Overall, teachers of both courses thought that the course that they taught provided a sound knowledge base for progression to university study. However, teachers of the traditional course were concerned that students do not acquire sufficient chemical knowledge when following a context led course, because the context course does not cover conceptual knowledge sufficiently. This was linked to the use of a spiral curriculum. However, teachers who taught the Salters course thought that the spiral approach was beneficial, because it allowed students the opportunity to revisit and revise topics, leading to greater understanding. Teachers using the concept led course thought that their chosen course could be taught in a logical sequence, but they did have concerns about continuity.

The Salters courses are designed to use more student-centred activities than traditional concept courses. Both groups of teachers thought that the Salters Chemistry courses were more student-centred, and used a wider range of teaching methods. Perhaps as a result, both groups of teachers thought that the Salters Chemistry course was interesting and motivating for students. Teachers who taught the Salters Chemistry course felt that it promoted good study skills and developed independent study in their students. In contrast, teachers who taught the traditional course were concerned that their students were too reliant on the textbook. The Salters course was considered to be more demanding to teach. This was largely due to the nature of the coursework at A2: an individual experimental investigation. Teachers found it challenging to manage large groups of students who were working on individual projects, both in terms of laboratory organisation and providing sufficient academic support to each student. However, the coursework was considered to be a useful learning activity for students. Teachers who taught traditional courses also thought that the Salters' coursework would be time-consuming. In an evaluation of the Salters Biology A level course, Lewis and Scott (2006) also found that teachers sometimes struggled to adjust to the more active learning approach used in the course at first, although this improved as they gained more experience. Furthermore, in an investigation of a Biology course with a shared assessment for concept and context led approaches, Braund et al. (2013) found that some teachers preferred to teach some topics using a context led approach, and other topics using a concept led approach.

The Salters A level Chemistry course was evaluated by Lubben and Bennett (2008) with respect to the Gilbert (2006) models. They concluded that the course was predominantly Model 2, with some elements of Model 3, because all examination questions were contextualised, and the supporting materials were organised by different contexts ("storylines"), which provided some opportunities for Model 3.

Chemie, Physik, Biologie im Kontext – Germany

The *Im Kontext* projects began in the late 1990s, and were initially based on the ideas and experiences resulting from the Salters project in England (Parchmann et al., 2006). The projects started as a result of national discussions in Germany about Germany's surprisingly weak performance in the TIMMS and PISA international comparison studies, leading to a general recognition that reform was needed at a national level (Parchmann et al., 2006). However, in Germany, each Bundesland (federal state) has a different school system, with a variety of different structures for Science education, leading to a wide range of different curricula. For example, in some Bundesländer, Science is taught as an integrated subject during early secondary education, while in others it is taught as three separate Sciences. This variety leads to significant challenges for the implementation of educational reform at the national level. The *Im Kontext* projects addressed this issue by using a symbiotic implementation strategy, in which teachers and researchers worked together in learning communities, to develop teaching units which were suitable for their own teaching situation. These units were then trialled by teachers in schools, and shared with other learning communities (Fey et al., 2004; Parchmann et al., 2006). While this approach has led to a feeling of ownership of the process by teachers, facilitating their own professional development, the lack of an overarching plan for the whole curriculum led to difficulties in providing systematically planned opportunities for students to transfer knowledge to other contexts. Furthermore, although students reported increased motivation, they felt that they sometimes got "lost in the context" (Pilot & Bulte, 2006). Additionally, teachers reported that they found it difficult to integrate a context led approach into existing curricula, and felt that they needed to place more emphasis on developing understanding of scientific concepts. However, this might have been the case because teachers spent time developing appropriate contexts, reducing their focus on concepts (Fey et al., 2004).

The *Im Kontext* projects value socially embedded group learning, which is promoted for both students following the courses, and those involved in the development process. As a result, Pilot and Bulte (2006) argue that the *Im Kontext* projects could be described as Model 4 under Gilbert's (2006) framework. However, given the autonomy with which the different learning communities operate, it is difficult to evaluate whether all curriculum units can be said to fall under the same model.

National Curriculum Statement – South Africa

Until 1995, the official South African curriculum (apartheid curriculum) was a very traditional, concept led curriculum, with little opportunity for

3. Salters Chemistry A level, offered by OCR, retained a separate assessment.

contextualisation. However, there were alternative curricula. In South African townships, the democratic movement promoted "People's Education", which valued students' life experiences, and provided opportunities for context based learning. Despite this, even within this movement, most contexts were provided as an addition to the scientific concepts. Between 1995 and 2006 an Interim Curriculum was introduced. Although curriculum documents mentioned the need for students to develop scientific literacy and prepare for the workplace, contexts were not used in the content specification, such that the Interim Curriculum could also be described as a concept led curriculum (Lubben & Bennett, 2008). Since 2006 the National Curriculum Statement has been used, which recognises the need to "value indigenous knowledge systems" (p.258), and to be able to use Science critically in various contexts. Textbooks developed to support this curriculum use context to exemplify concepts previously taught. However, some supplementary teaching activities do allow a greater interaction between context and concepts, as do some parts of the assessment, leading Lubben and Bennett (2008) to conclude that while the majority of the course could be described as Model 1, there are some elements of Models 2 and 3.

Chemistry in Context – USA

In the USA, university students study a broad curriculum, so that many Science departments teach students who are not planning to continue their study of Science, and who may or may not have studied particular areas in high school. As a result, some universities offer courses targeted at these students, recognising that they have different needs and interests from those who are planning to continue to study Science. *Chemistry in Context* is a university textbook aimed at students who are not planning to specialise in Chemistry at university (Schwartz, 2006). In this respect it differs from other context led Science courses discussed in this article, which aim to provide a foundation for further study as well as meeting the needs of students who will not continue to study Science. The textbook was developed by university teachers, on the basis of their own teaching experience, rather than educational research, and aims to motivate students to learn Chemistry, and understand its societal significance, while developing an understanding of the fundamental concepts of Chemistry. The concepts and contexts which are taught are organised on the principle of a spider's web, showing links between different concepts and contexts. The contexts which have been chosen are typically real-world societal problems; these contexts were chosen in preference to topics relating to students' self-interests, due to their maturity levels. However, the inclusion of such topics may be challenging to teach, because instructors are likely to be Chemistry specialists foremost, and may not have specialist knowledge of the societal issues included in the course. Only topics which had a significant chemical content were chosen, to allow students to develop their knowledge of Chemistry concepts. However, typically more information about the underlying Chemistry is provided than is needed to understand the context. Despite this, the selection of conceptual content was largely driven by the choice of contexts, because there was no need to cover particular content as a preparation for further study. This may help to prevent the curriculum becoming overloaded (Pilot & Bulte, 2006). Similar to the *Im Kontext* projects in Germany, the textbook has been used in different institutions working in different learning environments, so it is difficult to evaluate the impact of the course. However, Schwartz (2006) reports that students following courses using the textbook showed more positive attitudes towards the study of Chemistry. Pilot and Bulte (2006) estimate that the *Chemistry in Context* course supports a Model 3 or 4 approach, due to the importance of the context and the emphasis on active learning.

Discussion and implications for A level reform

Context led Science courses share the aim of making Science education more relevant to students' lives, increasing their interest in, and motivation to study Science. They are now used in many different educational contexts, and have been shown to be effective in increasing student motivation (Bennett et al., 2005; Braund et al., 2013; King, 2012; Parchmann et al., 2006; Schwartz, 2006). However, context led courses can also be characterised by their diversity. There are many different types of contexts which can be used as a framework to explain different scientific concepts, from issues which may directly impact on students' lives, to global issues which may have a less direct impact on their everyday lives. Alternatively, a context can serve to make students aware of ways in which Science is used in industry, which may increase their awareness of possible careers in Science. For a context to be used successfully, students must be able to engage with it, either at a personal level, or through an appreciation of the importance of an issue, and be able, with support, to make the link to the appropriate scientific concepts. The choice of context used in a Science course should therefore depend on the aims of the specific course, and the situation in which it is taught. However, there is a danger that students will spend too much time learning about the context, rather than the concept (Fey et al., 2004; Parchmann et al., 2006). Furthermore, if the structure of the curriculum does not allow students to revisit concepts in different contexts in a structured way, then they may not be able to transfer their understanding of a concept to a new context, nor develop a full understanding of that concept (King, 2012). For example, Barker and Millar (2000) found that students' experiences studying basic thermodynamics in the context of a fuel-oxygen system meant that they formed a strong association between covalent bond formation and energy release, which they found hard to extend to ionic bonding. However, as King (2012) notes, this is also true of traditional concept led courses.

Context led Science courses are typically associated with constructivist ideas surrounding teaching and learning, specifically, the need to draw on students' prior knowledge and understanding to allow them to engage actively in constructing meaning, so that learning can take place. The use of everyday contexts may help students to relate what they are learning to their everyday experiences. However, not all students are likely to be equally familiar with all contexts, and in the case of industrial contexts, very few students may have any direct experience of the contexts used. Furthermore, as noted by Peşman and Özdemir (2012), it is possible to use traditional transmission based pedagogies in a context led course, and student-centred pedagogies in a concept led course. However, since the supporting materials for many context led courses use a student-centred, active learning pedagogy, teachers who choose to use a context led course may need to learn new pedagogical skills, as well as developing their knowledge of the contexts to be taught in the course. This may make the introduction of context led courses particularly demanding for teachers (Lewis & Scott, 2006). As a result, the successful implementation of context led Science courses is dependent on the attitudes of the teachers and the support which they are given. Enabling teachers to contribute to the development of materials may help to develop their

sense of ownership of the projects, and contribute to the success of the context led approaches (Fey et al., 2004).

The perceived overloading of Science curricula was one motivation for the introduction of context led courses. However, it could be argued that adding context exacerbates this problem, by adding additional material to be taught, potentially at the expense of conceptual understanding. Gilbert (2006) proposes that the conceptual content should be reduced to make space for contexts. However, this is not always possible, particularly when the conceptual content to be taught is regulated. Bennett et al. (2005) found that teachers who taught concept led courses had concerns about conceptual development in context led courses. In general, little research suggests that students who follow a context led course are disadvantaged in terms of conceptual knowledge development (Braund et al., 2013; King, 2012). However, there are considerable challenges in comparing concept and context led courses. Firstly, it is not clear how concept knowledge should be assessed in a way that allows a direct comparison, because students from each group are used to answering questions framed in a different way: a context led student would presumably find it easier to answer a question framed in a context than a concept led student, and vice versa.

Implications for A level reform

Pilot and Bulte (2006) argue that the integration of the Salters courses with national, large scale assessments (e.g., A levels) was critical to its success. However, this creates a tension in those contexts where national assessments are heavily regulated, because it is necessary to design assessments which conform to regulatory requirements, while recognising that concept and context led courses need different assessment approaches. In England and Wales, reformed Science A levels will be first taught from 2015. Currently two exam boards offer context led Science A levels: OCR offers Salters Chemistry, and Advancing Physics (Ogborn, 2003), while Pearson-Edexcel offers Salters-Nuffield Biology and Salters-Horners Physics. Advancing Physics was originally developed with the Institute of Physics, although their financial interest in the course has now ended. Currently OCR offers different assessments for traditional A level courses and context led courses, whereas Pearson-Edexcel does not. When the reformed A levels are introduced, OCR will offer a full suite of context led A levels (Advancing Biology, Advancing Physics and Salters Chemistry), while Pearson-Edexcel will continue to offer context led A levels in Biology and Physics. Of these A levels, only the reformed Pearson-Edexcel Salters-Horners Physics course will use an assessment shared with a traditional concept led course.

An important feature of the Salters context led A levels is the individual experimental investigation. Teachers report that they consider the investigation to be educationally beneficial, though very difficult to manage, in terms of workload for themselves and their students (Bennett et al., 2005). The reformed A levels will share a framework for practical assessment: throughout their course, students will be required to conduct practical activities from twelve different areas. Although the framework is shared, within this model there is considerable scope for teachers to choose practical activities which match the type of course which they are teaching. For example, the context led OCR specification for Advancing Biology allows teachers to choose practical activities for each of the 12 areas (OCR, 2014). The inclusion of research skills as one of the 12 areas enables students to research appropriate contexts, and link these to the laboratory work which they undertake, allowing students to develop independent study skills throughout the course. This may reduce concerns which teachers have expressed about workload when teaching context led specifications, because high-stakes practical work will no longer be concentrated in one part of the course, potentially increasing uptake of the context led courses. However, this aspect of the reform will also reduce the distinctive nature of the context led A level courses, and reduce the scope for future innovation.

The choice of contexts used in a context led Science course is crucial to its success. It can be challenging to introduce contexts when the conceptual content is highly specified, as is the case for the reformed A levels, because there is a risk that the curriculum can become overloaded. However, the reform process provides an opportunity to reflect on the contexts used in a course, to ensure that the most appropriate contexts are used. For example, the new context led OCR A level in Biology (Advancing Biology) uses contexts which were selected using a variety of methods. Firstly, contexts which had been used in earlier context led courses were re-evaluated and updated, based on the experiences of teachers and developers. This is similar to the process used in the *Im Kontext* courses in Germany, where teachers were involved in the development and evaluation of context led materials. Secondly, as discussed above, there is a strong link between certain pedagogical approaches and context led teaching, and so course developers were also involved in the development of support materials for teachers. Both of these approaches help to ensure that the contexts which are chosen clearly highlight important biological concepts in a way which is clear to students and teachers. Additionally, the need to conform to content standards specified by the Department for Education led to particular emphasis being given to certain contexts and topics (e.g., natural selection), to ensure that the required conceptual content would be studied in sufficient depth. Finally, for those areas of the course which were new (e.g., Plant Biology), additional consideration was given to ensuring that contexts (such as food security) which are of particular contemporary importance were included, to help students to link their developing biological knowledge with issues which they may have encountered in the media. The assessment was developed to reflect these aims. For example, as part of the assessment, students will read a scientific article exploring a particular context, which will then be used as the basis for examination questions. When the course has been taught for the first time, further evaluation of the contexts chosen will be undertaken, based on the experiences of teachers.

Conclusions

The context led courses described in this article were developed either as a result of a top-down drive for reform (South Africa, Germany), or evolved in educational situations which allowed for diversity in the approach taken to Science teaching and learning (USA, England). Indeed, the German *Im Kontext* projects could be considered to be both, in that they were instigated at the national level, but developed to allow for diversity in different educational situations within different Bundesländer. When the Salters project began in England in the early 1980s, the regulatory frameworks in place allowed substantial diversity in assessment, so it was possible to develop Science courses which combined innovative approaches to teaching and assessment. Since then, however, increased regulation has led to much greater uniformity across qualifications. For the reformed Science A levels, the assessment requirements have been highly constrained, with common weightings for

the Mathematics (with variation across Science subjects), and a shared approach to practical assessment across awarding bodies. While this increased uniformity may lead to increased comparability across qualifications, it reduces the potential for important innovations such as the Salters project in the future.

References

Astin, C., Fisher, N., & Taylor, B. (2002). Finding physics in the real world: how to teach physics effectively with visits. *Physics Education*, *37*(1), 18.

Barker, V., & Millar, R. (2000). Students' reasoning about basic chemical thermodynamics and chemical bonding: what changes occur during a context-based post-16 chemistry course? *International Journal of Science Education*, *22*(11), 1171–1200. doi: 10.1080/09500690050166742

Bennett, J., Gräsel, C., Parchmann, I., & Waddington, D. (2005). Context-based and Conventional Approaches to Teaching Chemistry: Comparing teachers' views. *International Journal of Science Education*, *27*(13), 1521–1547. doi: 10.1080/09500690500153808

Bennett, J., & Lubben, F. (2006). Context_based Chemistry: The Salters approach. *International Journal of Science Education*, *28*(9), 999–1015. doi: 10.1080/09500690600702496

Bennett, J., Lubben, F., & Hampden-Thompson, G. (2013). Schools That Make a Difference to Post-Compulsory Uptake of Physical Science Subjects: Some comparative case studies in England. *International Journal of Science Education*, *35*(4), 663–689. doi: 10.1080/09500693.2011.641131

Braund, M., Bennett, J., Hampden-Thompson, G., & Main, G. (2013). *Teaching approach and success in A-level Biology: Comparing student attainment in context-based, concept-based and mixed approaches to teaching A-level Biology*. *Report to the Nuffield Foundation*. York: University of York, Department of Education.

Dunkerton, J. (2007). Biology outside the classroom: the SNAB visit/issue report. *Journal of Biological Education*, *41*(3), 102–106. doi: 10.1080/00219266.2007.9656077

Duranti, A., & Goodwin, C. (1992). *Rethinking context: Language as an interactive phenomenon*. Cambridge: Cambridge University Press.

Edexcel. (2014). GCE from 2008. Retrieved from http://www.edexcel.com/QUALS/GCE/GCE08/Pages/default.aspx

Fey, A., Gräsel, C., Puhl, T., & Parchmann, I. (2004). Implementation einer kontextorientierten Unterrichtskonzeption für den Chemieunterricht. *Unterrichtswissenschaft*, *32*(3), 238–256.

Giamellaro, M. (2014). Primary Contextualization of Science Learning through Immersion in Content-Rich Settings. *International Journal of Science Education*, *36*(17), 2848–2871. doi: 10.1080/09500693.2014.937787

Gilbert, J. K. (2006). On the Nature of "Context" in Chemical Education. *International Journal of Science Education*, *28*(9), 957–976. doi: 10.1080/09500690600702470

Gilbert, J. K., Bulte, A. M. W., & Pilot, A. (2011). Concept Development and Transfer in Context-Based Science Education. *International Journal of Science Education*, *33*(6), 817–837. doi: 10.1080/09500693.2010.493185

Institute of Physics. (2003). Personality: Keeping things in context – Liz Swinbank Teaching Anecdotes: The Wright Brothers Starting Out: What Katie did next: part 7. *Physics Education*, *38*(6), 536.

Kazeni, M., & Onwu, G. (2013). Comparative Effectiveness of Context-based and Traditional Approaches in Teaching Genetics: Student Views and Achievement. *African Journal of Research in Mathematics, Science and Technology Education*, *17*(1-02), 50–62. doi: 10.1080/10288457.2013.826970

King, D. T. (2012). New perspectives on context-based chemistry education: using a dialectical sociocultural approach to view teaching and learning. *Studies in Science Education*, *48*(1), 51–87. doi: 10.1080/03057267.2012.655037

King, D. T., & Ritchie, S. M. (2013). Academic Success in Context-Based Chemistry: Demonstrating fluid transitions between concepts and context. *International Journal of Science Education*, *35*(7), 1159–1182. doi: 10.1080/09500693.2013.774508

Lewis, J., & Scott, A. (2006). The importance of evaluation during curriculum development: the SNAB experience. *School Science Review*, *88*(323).

Lubben, F., & Bennett, J. (2008). From novel approach to mainstream policy? The impact of context-based approaches on chemistry teaching. *educación quimica,* 252.

Nawrath, D. (2010). *Kontextorientierung. Rekonstruktion einer fachdidaktischen Konzeption fur den Physikunterricht*. PhD Dissertation, Carl von Ossietzky Universitat Oldenburg.

OCR. (2014). *OCR Level 3 Advanced GCE in Biology B (Advancing Biology) (H422) Specification*. Retrieved from http://www.ocr.org.uk/Images/171714-specification-accredited-a-level-biology-b-advancing-biology-h422.pdf

Ogborn, J. (2003). Advancing Physics evaluated. *Physics Education*, *38*(4).

Overman, M., Vermunt, J. D., Meijer, P. C., Bulte, A. M. W., & Brekelmans, M. (2012). Textbook Questions in Context-Based and Traditional Chemistry Curricula Analysed from a Content Perspective and a Learning Activities Perspective. *International Journal of Science Education*, *35*(17), 2954–2978. doi: 10.1080/09500693.2012.680253

Overman, M., Vermunt, J. D., Meijer, P. C., Bulte, A. M. W., & Brekelmans, M. (2014). Students' Perceptions of Teaching in Context-based and Traditional Chemistry Classrooms: Comparing content, learning activities, and interpersonal perspectives. *International Journal of Science Education*, *36*(11), 1871–1901. doi: 10.1080/09500693.2013.880004

Parchmann, I., Gräsel, C., Baer, A., Nentwig, P., Demuth, R., & Ralle, B. (2006). "Chemie im Kontext": A symbiotic implementation of a context_based teaching and learning approach. *International Journal of Science Education*, *28*(9), 1041–1062. doi: 10.1080/09500690600702512

Peşman, H., & Özdemir, Ö. F. (2012). Approach–Method Interaction: The role of teaching method on the effect of context-based approach in physics instruction. *International Journal of Science Education*, *34*(14), 2127–2145. doi: 10.1080/09500693.2012.700530

Pilot, A., & Bulte, A. M. W. (2006). The Use of "Contexts" as a Challenge for the Chemistry Curriculum: Its successes and the need for further development and understanding. *International Journal of Science Education*, *28*(9), 1087–1112. doi: 10.1080/09500690600730737

Reiss, M. J. (2005). SNAB: a new advanced level biology course. *Journal of Biological Education*, *39*(2), 56–57. doi: 10.1080/00219266.2005.9655961

Reiss, M. J. (2008). The use of ethical frameworks by students following a new science course for 16–18 year-olds. *Science & Education*, *17*(8–9), 889–902. doi: 10.1007/s11191-006-9070-6

Schwartz, A. T. (2006). Contextualized Chemistry Education: The American experience. *International Journal of Science Education*, *28*(9), 977–998. doi: 10.1080/09500690600702488

Taasoobshirazi, G., & Carr, M. (2008). A review and critique of context-based physics instruction and assessment. *Educational Research Review*, *3*(2), 155–167. doi: http://dx.doi.org/10.1016/j.edurev.2008.01.002

Assessing active citizenship: An international perspective

Prerna Carroll, Simon Child and **Ellie Darlington** Research Division

Introduction

The evolution of citizenship studies in England and Wales

The introduction of citizenship as a formal part of the National Curriculum in 2002 was the result of years of momentum building through the publication of policy-steering documents, and the commonly held view that new generations of students were suffering from a lack of political engagement. The start of this movement was based on Marshall's (1950) influential work which argued that three elements of citizenship (civil, political and social) were developed in the eighteenth, nineteenth and twentieth centuries respectively. Citizenship was seen by Marshall as rights-based, with a large role of the state in ensuring that these rights are met across the three elements he identified.

However, a re-conceptualisation of citizenship in the UK occurred during the 1980s. Citizenship was being viewed as more than just the payment of taxes, but also the contribution of time and commitment (Orton, 2006). The Speaker's Commission of 1990 perceived two primary barriers to this more active participation in society. First, the report suggested that young people have little idea of their rights and responsibilities as citizens. Secondly, the report argued that citizenship has to be learned like any other subject, and that current provisions in schools were inadequate. The report recommended that citizenship education should be introduced across the curriculum and formally recorded. However, there was little detail offered as to how citizenship education in schools should be implemented, the target age group, or how assessment of learning and understanding should be structured.

These issues of implementation were addressed by the *Crick Report* in 1998. Crick (1998) had two main aims: to produce a statement of the aims and purposes of citizenship education in schools; and to provide a framework of what citizenship education may look like in schools. Following on from the *Speaker's Report*, Crick (1998) noted that the concept of 'active' citizenship was back in currency. The neo-liberal perspective underlying the definition of an appropriate citizenship education sees individuals as fully self-regulated, active members of the community, with little reliance on the state. This is in contrast to passive definitions of citizenship that place greater emphasis on status, national identity and obedience (Ross, 2008).

Perhaps ironically, Crick (1998) attached great importance to the role of formal, state-led education in developing individuals into self-regulated, active citizens. The report argued that citizenship education was "too important to be left to chance" (p.14) and recommended that "citizenship education is important and distinct enough to warrant a separate specification within the national framework" (p.18). It recommended that citizenship education should focus on three areas: social and moral responsibility; community involvement; and political literacy. Social and moral responsibility was defined as an understanding of the rule of law, concepts of fairness, and the environment. This was linked to community involvement, which was defined as the participation in activities that intend to serve others. Finally, political literacy was defined as not just knowledge of political institutions, but an understanding of how political decision-making is related to social or economic issues, and their solutions. The focus on political literacy was seen to be of particular importance given the perception that younger generations lacked engagement with the political process (see Miles, 2006, for a discussion).

The recommendations of the *Crick Report* were accepted by the UK government, and in 2002 citizenship education became part of the National Curriculum, two years after the introduction of the revised curriculum in other subjects. The National Curriculum (Department for Education and Skills [DfES], 2004) stated that by Key Stage 4 (KS4) students (age 16) should have acquired the following:

- Knowledge and understanding about becoming informed citizens;
- Developed skills of enquiry and communication; and
- Developed skills of participation and responsible action.

Assessing citizenship

The schools responsible for teaching citizenship are given a level of autonomy in how it is delivered. There are a variety of different approaches to its teaching (Boss, 2014), and GCSE Citizenship is one approach that has gained momentum in recent years. This qualification is competing with more formative or non-examined approaches adopted by some schools. Exam boards are required to assess students against three assessment objectives which test their ability to recall knowledge, apply skills and analyse and evaluate issues. Each board uses one or more assessment types to assess these skills, using a mix of internal and external assessment methods. For example, the Oxford, Cambridge and RSA (OCR) exam board assesses the unit 'Rights and Responsibilities – Getting Started as an Active Citizen' through a controlled assessment. Students are required to evaluate a citizenship campaign within their schools or community that promotes the rights and responsibilities of citizens (OCR, 2012).

In England and Wales, the exam boards are regulated by The Office of Qualifications and Examinations Regulation (Ofqual), a non-ministerial department of the UK government. In 2010, the government published a White Paper – *The Importance of Teaching* (Department for Education [DfE], 2010) – which outlined that qualifications should "match up to the best internationally in providing a good basis for [future] education and employment." (p.40). This resulted in a period of reform, with changes to both the National Curriculum and to the parameters guiding which qualifications would be accredited by the regulator. Changes included the movement to fully linear qualifications and the removal of internal assessment if a case could not be sufficiently made for its inclusion.

The draft curriculum for specific subjects was published in February

2013, and included details of what students should learn in citizenship at Key Stage 3 (KS3) and KS4. At KS4, for example, the National Curriculum states that pupils should be taught about the following:

- *Parliamentary democracy*
- *Electoral systems used in and beyond the United Kingdom*
- *Other systems and forms of government*
- *Local, regional and international governance*
- *United Kingdom's relations with the rest of Europe, the Commonwealth and the wider world*
- *Diverse national, regional, religious and ethnic identities in the United Kingdom*
- *Active participation in the community*
- *Wages, taxes, credit, debt, financial risk and a range of more sophisticated financial products and services*.

(DfE, 2013b)

In terms of assessment, it was determined that the reformed GCSE Citizenship will be assessed using external exam only, and that 25 per cent of the qualification will be based on the assessment of students' active citizenship.

GCSE Citizenship needs to meet the demands of the regulator, while simultaneously achieving the desired outcomes of a broad citizenship education. Ofqual's (2013) directive that external exams "should be the default method of assessment" (p.20) for reformed qualifications whenever possible presents a challenge to the formalised assessment of citizenship. Exam boards are required to articulate how desired, valid outcomes of citizenship education can be achieved through a dedicated qualification in the subject, and its constituent assessments. In particular, this is an issue for the assessment of 'active' citizenship, because it is underpinned by student participation and responsibility.

This tension is the focus of the current article which has three main aims. Firstly, this review aims to outline what is meant by active citizenship. Second, it aims to explore international approaches to the assessment of active citizenship to better understand how it is dealt with in other jurisdictions. Lastly, the review evaluates the different approaches to ascertain how active citizenship may be assessed by external exam.

Defining active citizenship

The first aim outlined above is to clarify what constitutes active citizenship. The term active citizenship is "a contested notion, imbued with different meanings and connotations" (Good Governance Learning Network [GGLN], 2013, p.12). It is a concept which is considered to be too country (and context) dependent to give a universal definition (Keser, Akar, & Yildrim, 2011; Menezes, 2003). It has roots in politics, and is often used "almost as a slogan that suits the politics of the day." (Kennedy, 2007, p.307). Nelson and Kerr (2006) describe active citizenship as being "fundamentally about engagement and participation" (p.iv). This engagement can be either "citizens engaging with the state" (electoral) or "citizens engaging with and among themselves" (civic) (GGLN, 2013, p.12; Annette, 2008).

Children's conceptions of active citizenship are shaped by their schooling, family, environment, the media and public figures (Crick, 1998). Hence, recommendations have been made for "practices oriented towards personal development, acquisition of social competencies for cohesion, integration and creativity." (Dimitrov & Boyadjieva, 2009, p.166). For example, children could become involved in the Junior Citizenship Programme, Community Service Volunteers, school councils or write to their local MP regarding issues which affect them (Crick, 1998) as ways of becoming active citizens through school.

Crick (2007), however, identified that active citizenship has two key components: action and knowledge. Crick argued that doing charitable work makes one a good citizen, but not an active one. An active citizen would also need the underlying knowledge behind why the social service was necessary. For example, children volunteering in a residential home would be deemed good citizens; however, active citizens would also understand the public policies, healthcare systems and personal circumstances that lead to the elderly being cared for in a residential home. Active citizens would be able to understand why volunteering was needed and even be able to suggest improvements and identify issues (Crick, 2007). Whilst Crick's definition of active citizenship is all encompassing, it is worth noting that it asks a lot of 16 year old learners. Perhaps the curriculum and assessment should provide them with the knowledge to enable them to develop into active citizens as they grow into adulthood, participating more in communities and taking on more social and civic responsibility?

What constitutes active citizenship appears ever-changing and greatly depends on context and country. However, the most common definitions stress civic and social responsibility coupled with knowledge and political literacy.

International perspectives on citizenship education

Citizenship is taught in several countries, each with its own interpretation of what constitutes being an active citizen. For this article, four education systems across five countries were studied: in England and Wales (treated together), the United States of America (USA), Australia and Singapore. Countries were chosen on the basis that citizenship was taught at secondary school, the syllabus included an element of active citizenship, and details on assessment were readily available through web searches or journal articles. All selected countries are economically developed, have established governments and have similar political contexts introduced in their syllabuses. Countries also use similar frameworks for citizenship education and assessment which focus on knowledge of the government policies and practices, economic and social issues, laws and rights and active citizenship. First, the development and structure of citizenship education and assessment in England and Wales is discussed followed by a review of practices in the USA, Australia and Singapore.

England and Wales

Due to citizenship being new to the National Curriculum in England and Wales in 2002, schools adopted a variety of different approaches to its incorporation as a subject (Kerr, Smith & Twine, 2008). One common approach was to incorporate citizenship education into related subject areas such as History, Geography and English (Crick, 1998; Ofsted, 2013). Keating, Kerr, Lopes, Featherstone, and Benton (2009) saw this approach as a barrier to effective citizenship learning, as students were often unaware of when they were being taught citizenship-related content. This view on the cross-curricular delivery of citizenship education is shared by

a recent report by Ofsted (2013). They suggested that, while some schools were confident they could deliver citizenship education without discrete provision, "the content was only partially relevant, often demonstrating little or no progression from KS3, and usually failed to fully meet objectives for citizenship" (p.25). In other words, there was evidence of an "implementation gap" (Kerr, Smith, & Twine, 2008, p.255) between the intentions of the *Crick Report* and how it is understood by teachers to inform their pedagogical approaches.

An alternative approach to the delivery of citizenship in schools is to work towards a GCSE qualification. This option was introduced by exam boards as a short course in 2002, before being extended to a full course option in 2008. There is some evidence to suggest that GCSE Citizenship is becoming an attractive option for schools, with increased entries in the full course GCSE option (Ofsted, 2013). Richardson (2010) found that teachers perceived summative assessment (specifically the GCSE) to be a useful tool to encourage students to take the study of citizenship seriously. She reported that students' motivation for subjects were typically underpinned by assessment, and that in schools where citizenship was not assessed, students would question its value.

The challenge for citizenship assessment (and qualifications more broadly) is to focus not just on knowledge, but also on how well that knowledge is understood, applied, debated and put into action outside the classroom (Quigley, 1995). These elements are central to achieving construct validity in citizenship assessment. In England and Wales the exam boards currently create a specification and assessment model that aims to examine students' learning outcomes against three Assessment Objectives (AOs):

> AO 1: Their ability to recall, select and communicate their knowledge and understanding of citizenship concepts, issues and terminology.
>
> AO 2: Their application of skills, knowledge and understanding when planning, taking and evaluating citizenship actions in a variety of contexts.
>
> AO 3: Their ability to analyse and evaluate issues and evidence including different viewpoints to construct reasoned arguments and draw conclusions.

For each AO, exam boards have one or more assessments. Currently each of the boards utilises a combination of controlled assessment and written tasks, with controlled assessment used to assess AOs that focus on active citizenship.

The formal assessment of citizenship has come under some criticism, as some concepts central to citizenship, such as 'active' citizenship, are difficult to define and thus to assess. Keating et al. (2009) found that teachers perceived difficulties with assessing active citizenship through controlled assessment at GCSE. It could be the case that while assessment in citizenship has the benefit of focusing the student's mind on the subject, it may encourage students to adopt surface learning approaches (Richardson, 2010).

United States of America

In the USA, there has been a push towards increasing citizenship studies, or civics education, since the education reform initiated by the current administration. As part of this reform, a 'road map' for civic education was developed in order to better inform students on civics, government, economics and history (State of Washington, 2014).

There are variations in how citizenship is taught within individual states. Internal and external assessment is used for different subjects and varies from state to state as well. In the State of Washington, civics education is taught throughout schooling and encourages the discussion of current local, national and international issues, and participation in school governance. Furthermore, it encourages schools to facilitate students' participation in community service linked to the formal curriculum as well as to engage them in extra-curricular activities in their community. In addition to this, students are also encouraged to take part in simulations of democratic procedures and processes such as voting, debates and elections. The subject is assessed internally by the teacher. Students are asked to prepare posters on a chosen topic and marks are based on students' ability to research, analyse, and evidence their knowledge.

In the State of Florida, civics includes similar content to the Washington curriculum. However, students must pass a Civics exam at the end of Grade 7 (children aged 12 to 13 years) in order to progress onto secondary school. The syllabus ensures that students have a good theoretical knowledge about the government and law and ensures that they learn about the "roles, rights, and responsibilities of United States citizens, and determine methods of active participation in society, government, and the political system" (Seminole County Public Schools, 2013, p.3). Students take an external exam in the form of a multiple-choice exam that tests all aspects of the Civics curriculum such as Geography and History. Aspects of active citizenship are assessed through questions that put the student in a hypothetical situation such as asking students how they would encourage their communities to provide low cost flu vaccinations (Florida Virtual School, 2014).

In order to ensure standardisation, the National Assessment of Educational Progress (NAEP), a research based division of the Department of Education, periodically assess a sample of students across the country on many subjects, including civics. The assessments are developed according to a quality framework and measures are taken to ensure reliability of scores (NAEP, 2014). The NAEP design Civics assessments based on five content areas:

1. *What are civic life, politics, and government?*
2. *What are the foundations of the American political system?*
3. *How does the government established by the Constitution embody the purposes, values and principles of American democracy?*
4. *What is the relationship of the United States to other nations and to world affairs?*
5. *What are the roles of citizens in American democracy?*

(NAEP, 2011, online)

The fifth content area appears most related to active citizenship as it directly places importance on the responsibilities of citizens as members of their society. In 2010, 21 per cent of the NAEP Civics assessment was dedicated to the roles of citizens and occurred through a range of question types, such as multiple-choice (MCQ), short response and extended response questions. Whilst the multiple-choice and short response questions were similar in nature to those found in the Florida exams (Florida Virtual School, 2014), the extended response questions enabled students to discuss, debate and rationalise their knowledge in a simulated context. One question, for example, from a past test required students to look at charts related to volunteering activities and asked what motivates people to volunteer. Based on the information provided,

students were then asked to choose three types of volunteer activity and to "identify specific actions" individuals can take outside their homes and explain "how it will make a difference in their own community." (NAEP Question Tool, 2010, online). This question required students to not only discuss ways in which people can volunteer (action), but also to deliberate the merits and consequences of volunteering (knowledge). This aspect of critical thinking and evaluation can often be missed when assessing students through practical work alone (Crick, 2007).

Australia

In Australia, the Department for Education, Science and Training (DEST) developed the Discovering Democracy curriculum and teaching materials to be taught in primary and middle schools across Australia in 1997. Since then, schools have incorporated this curriculum into their schooling; however, the interpretation and implementation of the syllabus is varied (Print, 2008). The latest reforms on the Civics and Citizenship curriculum have been set out by the Australian Curriculum, Assessment and Reporting Authority (ACARA) and describe a curriculum split into two interrelated strands; 'Knowledge and Understanding' and 'Skills' (ACARA, 2014, online). The new curriculum is implemented in the curriculum from Year 3 to Year 10. For the strand 'Knowledge and Understanding', students focus on three areas at each year level; Government and Democracy, Laws and Citizens and Citizenship, Diversity and Identity. For 'Skills', students develop knowledge of Questioning and Research and Problem Solving and Decision making (ACARA, 2014, online).

At Year 9 and 10, students are assessed on their ability to evaluate, assess and critically analyse features of the Australian political and legal systems. All assessment in this course, and other subjects in Australia, is marked and reviewed by teachers. However, as a result, student outcomes vary significantly. According to test data from the Ministerial Council for Employment, Education, Training and Youth Affairs (MCEETYA, 2006), students in Years 6 and 10 know relatively little about the political system and citizenship in Australia. This finding could be due to schools not fully or systematically introducing this curriculum into their school system.

Similar to the NAEP assessments in USA, the Australian Curriculum Assessment and Reporting Authority (ACARA), a statutory authority responsible for the management and development of the National Curriculum (similar to Ofqual in England), regularly sample Year 6 and Year 10 students on a range of subjects on a rolling three yearly basis (National Assessment Program [NAP], 2010a). The Civic and Citizenship test covers topics such as the historical and current policies and government practices, laws, rights and responsibilities, and local, regional and global influences on Australian economy. The tests are delivered online and include a range of multiple-choice and short answer questions. Questions related to active citizenship tend to present a situation and ask the student to rationalise or reason for or against certain behaviours. In addition to the test, students are asked to complete a questionnaire about their extracurricular and wider volunteering activities. Similar to the NAEP tests, these questions go beyond simply recognising what constitutes being a good citizen and require students to rationalise and justify the principles behind the actions.

Singapore

In Singapore, students in local secondary schools have Character and Citizenship Education (CCE) as a mandatory subject in their curriculum. According to the latest syllabus published by the Ministry of Education (MOE), the goals of the course are to instil key values and competencies in students that enable them to be "good individuals and useful citizens" (MOE, 2014, p.1). The syllabus is made up of three components – "Core values", "Social and emotional competencies" and "Citizenship" – and takes up 60 hours per year (MOE, 2014, p.1). The citizenship component of the syllabus appears to be the most closely linked to GCSE Citizenship course and its key components are:

- Active community life
- National and cultural identity
- Global awareness
- Socio-cultural sensitivity and awareness

The CCE syllabus has been carefully developed based on cognitive constructivist theory and focusses on the students' perspective on learning. The constructivist theory of learning proposes that teachers cannot force knowledge on students. Instead, students construct their understanding from their daily experiences and social interactions with others (Nucci & Narvaez, 2008). These experiences then enable students to process new information and modify their current understanding accordingly (Strommen & Lincoln, 1992). As a result, suggested teaching methods emphasise developing skills and internalising values through action and reflection where the end result is "something more meaningful other than a grade" (MOE, 2014, p.39). Suggested teaching methods include storytelling, role-playing, dialoguing and group work.

The syllabus uses internal assessment models including self-assessment, peer assessment and teachers' assessment. Assessments could range from research projects, posters and/or debates. Unlike most other qualifications in Singapore, there is no external assessment for this course as it is designed to holistically develop the students. Students are expected to self- and peer-assess so they can reflect on their own performance and knowledge. However, a purely internal approach can pose issues as assessment is wholly dependent on teachers' observations and, in cases where peer assessment is used, it could be prone to bias. Whilst the curriculum may encourage self-learning and development, the assessment method may have some disadvantages. Internal assessment models, however, can test a wider range of skills that cannot be tested by external written assessments.

Discussion and implications

This study aimed to explore the conceptualisation and assessment of active citizenship from several international perspectives. The aim of the research was to identify models of assessment that validly and reliably test the skills and understanding that underlies active participation. This is in reaction to the educational reforms currently underway in England and Wales, where all GCSEs are undergoing substantial changes which include changes in subject content, difficulty and assessment (DfE, 2013; Ofqual, 2013). As part of this change, GCSE Citizenship is being reformed to be completely externally assessed, where previously 25 per cent of the course was internally assessed. Exam boards have to ensure that the new course meets the demands of the regulator and ensure that the desired outcomes of the course are met. One such learning outcome is to ensure that students who complete the course are active citizens in their community. However, this skill has previously always been assessed internally via controlled assessment. It was an aim of this research to

define what constitutes active citizenship and, using international perspectives, identify models of assessment that validly and reliably test these skills.

Active citizenship was defined in this study as an amalgamation of knowledge (political literacy) and action (civic duty) (Annette, 2008; Crick, 2007). As such, an assessment which tests both these constructs would be needed to provide a valid measure of active citizenship. Four education systems across five countries were investigated as part of this study: England and Wales; the USA; Australia; and Singapore. We found that internal models of assessment were largely favoured when teaching active citizenship. External assessment was used as a measurement tool to determine progress and standards of education nationally (NAEP and NAP tests in USA and Australia respectively) or in order to progress to further education (Florida).

Following the analysis of the different types of assessments used in the selected countries, it was possible to identify models of assessment best suited to assessing active citizenship (see Table 1). This includes both internal assessments, which is the focus of most assessment approaches taken by different jurisdictions, and external assessments, the preferred mode of assessment in the UK's most recent educational reform (Ofqual, 2013). There are many factors that determine the validity and reliability of assessment. A key concern when considering validity in assessment is to ensure that an assessment measures the skills it is intended to measure. As such, any assessment that can measure 'action' and 'knowledge' in citizenship would contribute to ensuring the assessment was valid. Reliability refers to comparability and consistency of the assessment. It aims to ensure that comparisons can be made between students' achievement and achievement over time (Jones & Bray, 1986). There are many factors that affect reliability, such as human factors and objectivity. Internal assessment of coursework could be prone to the same level of bias (e.g., tiredness of the examiner) or objectivity as an externally assessed extended response question. As such, measures need to be put in place to ensure that mark schemes and moderation practices are robust to increase reliability of assessment outcomes in either context.

Table 1: Types of assessment used to assess active citizenship internationally (UK, USA, Australia and Singapore)

Types of assessment	*Key skills for active citizenship*	
	Knowledge	*Action*
Internal Assessment		
Research project[1]	✓	✓
Report writing[1,3]	✓	✓
Community Service[1,2,3]	✓	✓
Simulations of democratic procedures[1,2,3]	✓	✓
Debates[2,4]	✓	✓
Speeches[2,4]	✓	✓
Posters[2,3,4]	✓	✓
External Assessment		
MCQ[2,3]	✓	
Short answer[1,2,3]	✓	
Extended response[1,2]	✓	✓
Hypothetical situation[1,2]	✓	✓
Personal case studies[1,2]	✓	✓

1. England and Wales; 2. USA; 3. Australia; 4. Singapore

Tick marks on the table indicate areas where this type of assessment, stimulus or question would be able to address the skills required when assessing active citizenship. The most common definitions of active citizenship stress the importance of action based on underlying knowledge and political literacy (Crick, 2007). Knowledge and action can be tested through all the internal assessment methods identified in this review. However, by evaluating current practice in a number of countries, extended response questions appear to be the external assessment method most likely to facilitate an appropriate assessment of active citizenship. The extended response questions required students to identify actions that defined a good citizen and discuss the underlying socio-political issues. These responses seem the most suitable as they require students to identify action and demonstrate their knowledge.

Whilst reflecting on activities they have conducted over the school year (such as volunteer work) in the extended response question would be ideal, ensuring reliability of scores across students would be challenging. Students from different socio-economic backgrounds, schoolings and communities could have very different experiences and therefore standardising marks based on those would provide an additional challenge. Furthermore, there could be issues with providing evidence that the students were actually involved in such activities. An extended response question providing a hypothetical context may alleviate the differences between pupils and remove the issue of asking students to evidence their active citizenship.

There are several implications from this research. Firstly, assessing active citizenship, as defined by this study, would require measuring students' ability to engage in civic duty and responsibility as well as their underlying knowledge of socio-political and economic issues. Secondly, internal assessment (e.g., a task administered by a teacher) is a common way to assess citizenship in other countries, and would appear to have advantages in that students can actively engage in the community and explore why their actions are necessary. Lastly, extended response questions, as used internationally, appear to be an appropriate method of testing active citizenship through external assessment. Whilst it does not guarantee that students are actively participating in the community, it ensures that students know what constitutes active participation and can, at the very least, simulate active citizenship. Further research could attempt to establish the validity of the different assessment methods, both as a measure of active citizenship within the qualification as a predictive measure (i.e., do students proceed to become active citizens in the future?).

References

ACARA (2014). *Civics and Citizenship*. Retrieved from http://www.australiancurriculum.edu.au/humanities-and-social-sciences/civics-and-citizenship/content-structure

Ahmed, A., & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy and Practice*, *18*(3), 259–278.

Annette, J. (2008). Community involvement, civic engagement & service learning. In J. Arthur, I. Davies, & C. Hahn, *The SAGE Handbook of Education for Citizenship and Democracy* (pp.388–398). London: SAGE Publications Ltd.

Black, B., Suto, I., & Bramley, T. (2011). The interrelations of features of questions, mark schemes and examinee responses and their impact upon marker agreement. *Assessment in Education: Principles, Policy and Practice*, *18*(3), 295–318.

Boss, G. (2014). *The role and purpose of Citizenship in the curriculum*. Retrieved from http://www.citized.info/pdf/students/George%20Boss.pdf

Crick, B. R. (1998). *Education for citizenship and the teaching of democracy in schools*. London: Qualifications & Curriculum Authority (QCA).

Crick, B. R. (2007). Citizenship: The political and the democratic. *British Journal of Educational Studies*, *55*(3), 235–248.

DfE. (2010). *The Importance of Teaching – The Schools White Paper*. Retrieved from https://www.gov.uk/government/publications/the-importance-of-teaching-the-schools-white-paper-

DfE. (2013a). *GCSE subject content and assessment objectives*. Retrieved from https://www.gov.uk/government/consultations/gcse-subject-content-and-assessment-objectives

DfE. (2013b). *National Curriculum in England: citizenship programmes of study for key stages 3 and 4*. Retrieved from https://www.gov.uk/government/publications/national-curriculum-in-england-citizenship-programmes-of-study/national-curriculum-in-england-citizenship-programmes-of-study-for-key-stages-3-and-4

Dimitrov, G., & Boyadjieva, P. (2009). Citizenship education as an instrument for strengthening the state's supremacy: An apparent paradox? *Citizenship Studies*, *13*(2), 153–169.

Florida Virtual School. (2014). *MJ Civics End-of-Course Practice Exam*. Retrieved from http://www.flvs.net/areas/studentservices/EOC/Documents/Civics%20Practice%20Test.pdf

Good Governance Learning Network (GGLN). (2013). *Active Citizenship matters: Perspectives from civil society on local governance in South Africa*. Cape Town: Harlen, W. (2007). Assessment of learning. London: SAGE Publications Ltd.

Jones, R.L. & Bray, E. (1986) *Assessment: From principles to action*. London: Macmillan Education Ltd.

Keating, A., Kerr, D., Lopes, J., Featherstone, G., & Benton , T. (2009). *Embedding citizenship education in secondary schools in England (2002–08)*. London: National Foundation for Educational Research.

Kennedy, K. J. (2007). Student constructions of 'active citizenship': What does participation mean to students? *British Journal of Educational Studies*, *55*, 304–324.

Kerr, D., Smith, A., & Twine, C. (2008). Citizenship education in the UK. In J. Arthur, I. Davies, & C. Hahn, *The SAGE Handbook of Education for Citizenship and Democracy*. (pp.252–262). London: SAGE Publications Ltd.

Keser, F. A. (2011). The role of extracurricular activities in active citizenship education. *Journal of Curriculum Studies*, *43*(6), 809–837.

Marshall, T. H. (1950). *Citizenship and social class and other essays*. Cambridge: Cambridge University Press.

Menezes, I. (2003). Participation experiences and civic concepts, attitudes and engagement: implications for citizenship education projects. *European Educational Research Journal*, *2*(3), 430–445.

Ministerial Council for Employment, Education, Training and Youth Affairs (MCEETYA). (2006). *National Assessment Program- Civics and citizenship Years 6 and 10 report*. Canberra: MCEETYA.

Ministry of Education. (2014). *Character and citizenship education: secondary*. Singapore: Student Development Curriculum Division.

Morgan, C. (1996). The teacher as examiner: the case of mathematics coursework. *Assessment in Education: Principles, Policy and Practice*, *3*(3), 353–375.

NAEP. (2010). *NAEP questions tool*. Retrieved from http://nces.ed.gov/nationsreportcard/ITMRLSX/

National Assessment Program (NAP). (2010a). *The tests*. Retrieved from http://www.nap.edu.au/nap-sample-assessments/napsa-the-tests.html

Nelson, J., & Kerr, D. (2006). *Active citizenship in INCA countries: Definitions, policies, practices and outcomes*. QCA.

Nucci, L. P., & Narvaez, D. (2008). *Handbook of Moral and Character Education*. UK: Routledge.

OCR (2012). *GCSE 2012 Citizenship Studies Specification*. Retrieved from http://www.ocr.org.uk/Images/82006-specification.pdf

Ofqual. (2013). *GCSE reform consultation June 2013*. Retrieved from http://www.ofqual.gov.uk/files/2013-06-11-gcse-reform-consultation-june-2013.pdf

Ofsted. (2013). *Citizenship consolidated? A survey of citizenship in schools between 2009 and 2012*. London: Office for Standards in Education, Children's Services and Skills.

Orton, M. (2006). Wealth, citizenship and responsibility: The views of "better off" citizens in the UK. *Citizenship Studies*, *10*(2), 251–265.

Quigley, C. B. (1995). *Issues concerning a national assessment of civics. National Assessment of Educational Progress*. Washington DC: Center for Civic Education.

Richardson, M. (2010). Assessing the assessment of citizenship. *Research Papers in Education*, *24*(4), 457–478.

An investigation into the numbers and characteristics of candidates with incomplete entries at AS/A level

Carmen Vidal Rodeiro Research Division

Introduction

AS and A levels are the most popular qualifications taken by students between the age of 16 and 18 in England. A levels are usually spaced out over two years and are made up of two types of units: AS units and A2 units. Since 2000, AS units can be supplemented by A2 units to complete a full A level qualification or they can be a qualification in their own right.

The existing AS qualification has allowed students to study a wide range of subjects and in some instances has meant students have taken subjects at A level in which they were not previously particularly interested and otherwise might not have pursued. Also, the AS levels in their current form are valued by universities and can encourage pupils from disadvantaged backgrounds to continue their studies (Watson, 2013).

Students normally take four subjects at AS level and then continue to study only three at A level. But, how do they decide which subjects to pursue at a higher level and which one to drop?

Sharp (1996) found that students who drop a subject do so for a

number of reasons and it is difficult to judge which ones are the most influential. These reasons include employment-related ones, organisation and content of the course, liking of the teacher, lack of enjoyment, lack of perceived usefulness or considerations of ability and difficulty. A research study by Pinot de Moira (2002) showed evidence that students who dropped subjects from AS level to A level usually had a bad result in the AS part of the examination.

In a survey of over 6500 AS/A level students carried out in 2006, Vidal Rodeiro (2007) found that Modern Foreign Languages were among the most dropped A level subjects, together with General Studies, Further Mathematics and Applied Information & Communications Technology (ICT). These subjects at AS level were probably used to encourage study in a breadth of areas, with the aim of broadening students' educational experience, or to allow students to study a subject in which they had an interest or skill outside of their core A level subjects ('core' meaning those they would like to pursue further, for example in Higher Education). The least dropped subjects were in the Creative Arts and Humanities fields.

In a study investigating the uptake of AS levels from 2007 to 2013, Sutch (2014) found that most students choose AS subjects from a range of subject domains[1] and that only around 14 per cent of students confine all their AS levels to just one. This recent research also showed that Modern Foreign Languages were still among the most dropped subjects (each dropped by around a third of students) together with Critical Thinking, General Studies and Citizenship. Furthermore, the study revealed that dropping rates have been increasing over time, particularly for Mathematics, Further Mathematics and Science subjects, and that they differed considerably by gender and academic ability.

In 2012, proposals for a reform of AS and A level qualifications were published (Gove, 2012). The proposals arose as a result of the concerns outlined in the Government's Education White Paper *The Importance of Teaching* (Department for Education [DfE], 2010) regarding A levels not being a good preparation for undergraduate study. The proposals were: for universities to be involved in the design and development of A levels; to consider whether the division of A levels into AS and A2 should continue; and to consider whether January re-sits should be allowed.

In 2013, the DfE (Gove, 2013) announced that the AS level would be a standalone qualification, at the same level as the A level, rather than as part of an A level. Separating both qualifications means that students will be able, if they want, to take new A levels without also taking an AS in the subject (if students take an A level after doing the AS, they will be reassessed on the material they have already covered). However, concerns have been raised about this move, as without a direct link to the A levels, the new style AS levels may not be as beneficial. For example, the Independent Schools Council warned that the reform of the AS qualification could reduce participation in harder subjects such as Mathematics and Languages (Stewart, 2013). Furthermore, the University of Cambridge has voiced strong criticism of the changes to AS levels, issuing a statement saying that they will "jeopardise over a decade's progress towards fairer access to the University of Cambridge." (BBC News, 2013). Similarly, an Oxford Admissions Tutor, speaking at a Westminster Education Forum seminar, said that "… the decoupling of AS levels from A levels will make students from disadvantaged backgrounds less likely to progress to university".

With the AS and A level qualifications reform in mind, the main aim of this article is to gain an understanding of the numbers and types of students who start but do not complete their AS and A level qualifications. This could help to anticipate changes in the uptake of the new AS levels.

In particular, this research addresses the following questions:

- After attempting some AS/A2 units, how many candidates drop out before achieving an AS or A level qualification?
- How does the performance of candidates who drop out before certificating at AS or A level compare to the performance of those who continue and certificate in the qualifications?
- Which types of students are more likely to drop out from AS to A level?

AS and A level incomplete entries were investigated in the following three subjects: Biology, Psychology and English Literature. Those subjects have been among the first ones to be reformed and new specifications will be in schools for first teaching in September 2015 (Ofqual, 2014).

The next section provides a description of the data and methods used in the research. The outcomes of the analyses are then presented and the final section brings all the results together and draws some conclusions.

Data and methods

Data

Details of awards in the Oxford, Cambridge and RSA (OCR) AS and A level qualifications in the two-year period leading to June 2013 were obtained from OCR's examination processing system. This data comprised student details (gender, date of birth and school) and assessment details (units, sessions, unit marks, unit grades, unit predicted grades and overall grades).

The focus was on 'typical' A level candidates who were at the end of Key Stage 5 (KS5) in the academic year 2012/13. Those candidates would have had to certificate for AS and/or A level qualifications in the typical four sessions up to the end of KS5 (January 2012, June 2012, January 2013, June 2013). Note that unit and overall re-sits were removed from the data (where candidates re-sat an examination, only the highest grade was kept).

This research also used data from the 2011 Key Stage 4 (KS4) and the 2013 KS5 extracts of the National Pupil Database (NPD)[2]. Students' characteristics such as previous performance at GCSE, AS subjects studied and type of school attended, were obtained from the NPD extracts and subsequently matched to the OCR data.

For the analyses carried out in this research, schools were classified as independent, selective, state-maintained (academies and comprehensive schools), sixth form colleges and further education (FE) colleges.

It should be noted that the matching between students who sat units in OCR specifications and students in the NPD was attempted using a Unique Pupil Number (UPN) common in both databases. However, in the OCR data there were students who did not have a UPN assigned to them and therefore a match (if indeed it existed) could not be found. This restricted the numbers of students available in some of the analyses.

1. Arts, English, Languages, Science/Mathematics, Social Science/Humanities.
2. The NPD, which is compiled by the Department for Education, is a longitudinal database for all children in schools in England, linking student characteristics to school and college learning aims and attainment. In particular, it holds student and school characteristics such as age, gender, ethnicity, level of deprivation, attendance and exclusions, matched to student level attainment data (Key Stage 2 to Key Stage 5 assessments). Students who start in a school/college are only recorded on the NPD if they enter for a qualification; those who leave school/college after a short time or do not sit examinations are not present in the data.

Method

The research questions were mainly addressed using descriptive statistical analyses. However, in order to identify the types of students that were most likely to drop out from AS to A level, multilevel logistic regression models were also employed.

Logistic regression is a type of regression analysis that is used when the dependent variable or outcome is a dichotomous variable (i.e., it takes only two values, which usually represent the occurrence or non-occurrence of some event) and the independent variables are continuous, categorical, or both. It is used to model the probability that the event of interest will occur as a function of the independent variables (see, for example, Hosmer & Lemeshow, 2000). A multilevel model was proposed due to the hierarchical (or multilevel) structure of the data. If we failed to recognise this hierarchical structure, then the standard errors of the regression coefficients would be underestimated, leading to an overstatement of the statistical significance. Detailed discussions of the implementation and outcomes of multilevel logistic regression models can be found in Snijders and Bosker (1999) or Goldstein (2011).

For the purpose of the analyses presented in this article, the dependent variable is 'drop out from AS to A level[3]', and the regression models take the following form:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 IV1_{ij} + \beta_2 IV2_{ij} + \cdots + \beta_l IVl_{ij} + u_j + \mathrm{e}_{ij}$$

where p_{ij} is the probability of student i in school j dropping out, $IV1$ to IVl are the independent variables, β_0 to β_l are the regression coefficients, u_j is a random variable at the school level and e_{ij} is an individual level residual. A detailed breakdown of the dependent and independent variables included in the multilevel logistic models is presented in Table 1.

Results

The OCR AS/A level Biology specifications (H021/H421) are unitised specifications. Each student must take three AS units, normally in the first year of study, to certificate for an AS level and then three A2 units for certification at A level.

Similarly, the OCR AS/A level Psychology specifications (H168/H568) and the OCR AS/A level English Literature specifications (H071/H471) are also unitised. However, in each of these two subjects, a student must take two AS units, normally in the first year of study, to certificate for an AS level and then two A2 units for certification at A level.

Up to June 2013, students were able to take different units in different sessions (January and June). From 2014, both AS units and A2 units are assessed in June only. A brief description of the units is given in Table 2. Further details about these specifications can be found in OCR (2013a; 2013b; 2013c).

Table 3 shows, for each subject, the numbers and percentages of candidates (among 'typical' ones) who sat at least one unit in the sessions from January 2012 to June 2013 and who certificated for an AS/A level qualification or dropped out after attempting some AS or A2 units.

Around half of the candidates certificated for both an AS and an A level qualification in Biology and Psychology, whilst over 40% certificated for an AS level only. Less than 1% of the candidates obtained an AS level and attempted some A2 units but dropped the subject before achieving the full A level. In English Literature, over 60% of the candidates certificated for both an AS and an A level qualification, whilst just under 27%

3. Having certificated for an AS level.

Table 1: Description of the variables included in the multilevel logistic regression models

Name	*Description*	*Range of values*
Dependent Variable		
Drop out from AS to A level	Indicator of dropping out from AS to A level (having certificated for an AS level)	Discrete variable: 0 did not drop out; 1 dropped out
Independent Variables		
Gender	Gender of the candidate	Discrete variable: male; female
GCSE subject	Indicator of whether the subject was taken at GCSE or not[a]	Discrete variable: 0 did not take the subject; 1 took the subject
Grade in GCSE subject	Grade achieved in the GCSE subject[b]	Discrete variable: A*, A; B or below[c]
Type of school	Type of institution the candidate obtained the AS/A levels in	Discrete variable: state-maintained; independent; sixth form college; selective; FE college
Average GCSE score	Average grade across all GCSE subjects taken	Continuous variable: real values in the range 0 to 8 (inclusive)
AS level grade	Grade achieved in the AS level qualification	Discrete variable: A; B; C; D; E; U
Number of AS subjects	Number of subjects attempted at AS	Continuous variable: its range depends on the subject

a. For Biology, it will be the type of Science taken at GCSE (Biology versus Additional Science).
b. This variable was only included in the models for Biology, as Science is compulsory at GCSE level. Psychology is not compulsory at GCSE and the majority of the students included in the analyses did not obtained a GCSE in it. English Literature, although not compulsory, is usually taken by around 70 per cent of the cohort.
c. There is hardly any progression to A level from candidates with grades below C at GCSE (e.g., Sutch, 2013).

Table 2: Overview of the OCR AS/A level specifications considered in this research

Subject	*Unit*	*Type of unit*	*Type of assessment*	*Weight*	*Maximum Uniform Mark Scale (UMS)*
Biology (H021/H421)	F211	AS	Written paper	30% (AS) - 15% (A)	90
	F212	AS	Written paper	50% (AS) - 25% (A)	150
	F213	AS	Coursework	20% (AS) - 10% (A)	60
	F214	A2	Written paper	15% (A)	90
	F215	A2	Written paper	25% (A)	150
	F216	A2	Coursework	10% (A)	60
Psychology (H168/H568)	G541	AS	Written paper	30% (AS) - 15% (A)	60
	G542	AS	Written paper	70% (AS) - 35% (A)	140
	G543	A2	Written paper	25% (A)	100
	G544	A2	Written paper	25% (A)	100
English Literature (H071/H471)	F661	AS	Written paper	60% (AS) - 30% (A)	120
	F662	AS	Coursework	40% (AS) - 20% (A)	80
	F663	A2	Written paper	30% (A)	120
	F664	A2	Coursework	20% (A)	80

Table 3: Candidates with at least one unit in the period of study, by type of qualification obtained

Units/Qualifications	*Biology*		*Psychology*		*English Literature*	
	Number of candidates	*%*	*Number of candidates*	*%*	*Number of candidates*	*%*
AS units only	1,826	5.49	633	3.85	233	1.69
AS qualification only	13,370	40.23	7,544	45.89	3,705	26.95
AS qualification + A2 units	259	0.78	121	0.74	8	0.06
AS and A level qualifications	16,435	49.45	7,868	47.86	8,524	62.00
Not AS but A level	1,343	4.04	272	1.65	1,279	9.30
Total	**33,233**		**16,438**		**13,749**	

certificated for an AS level only. Only eight candidates obtained an AS level and attempted some A2 units but dropped the subject before achieving the full A level. Finally, around 6% of the candidates in Biology, 4% in Psychology and 2% in English Literature did not achieve any qualification and dropped out after attempting at least one AS unit.

Table 3 also shows that there were some candidates (approximately 4% in Biology, 2% in Psychology and 9% in English Literature) who had an A level result but not an AS result. Some of these candidates might have aggregated for the AS level prior to January 2012 and some of them might have aggregated towards an A level only (although they might have had enough units to certificate for an AS level as well). It should be noted that to obtain an A level, candidates do not need to have been entered for the AS level first (OCR, 2013d).

In Biology, more than half of the students who dropped out before achieving the AS qualification attempted only one unit (60%). However, there was a reasonably large percentage of candidates (35%) who attempted three units, enough for AS certification, but decided not to aggregate. The average Uniform Mark Scale (UMS) percentage in the AS units for these candidates was 49%, which would have led to a grade E at AS[4]. Similarly, over 80% of the candidates who certificated for the AS level in Biology, but did not achieve an A level in the subject, only attempted one A2 unit and just over 8% attempted either two or three units.

A similar pattern emerged in Psychology, where the majority of the candidates who dropped out before achieving the AS qualification attempted only one unit (82%). As above, there was also a reasonable percentage of candidates (18%) who attempted two units, enough for AS certification, but decided not to aggregate. The average UMS percentage in the AS units for these candidates was 44%, which would have led, again, to a grade E at AS. Similarly, only two candidates who certificated for an AS level in Psychology but did not achieve an A level in the subject attempted two A2 units; the remaining 119 candidates attempted just one A2 unit.

Surprisingly, and contrary to the patterns for Biology and Psychology, the majority of the candidates who dropped out before achieving the AS level qualification in English Literature attempted two units (85%), enough for AS certification, but decided not to aggregate. The average UMS percentage in the AS units for these candidates was 67%, which would have led to a grade C at AS. Regarding the number of A2 units attempted by candidates who certificated for an AS level in English Literature but did not achieve an A level in the subject, three out of the eight candidates in this group attempted two A2 units; the other five candidates attempted just one A2 unit.

Tables 4 and 5 present the performance (in AS and A2 units, respectively) of candidates who dropped out before certificating at AS or A level and compare that to the performance of those who continued and certificated in the qualifications.

These tables show that the average unit performance, in terms of the UMS percentage achieved, increased with the increasing level of the qualification.

Table 4: Average UMS percentage in AS units, by type of qualification obtained

Units/ Qualifications	*Biology*		*Psychology*		*English Literature*	
	Average UMS %	*Standard Deviation*	*Average UMS %*	*Standard Deviation*	*Average UMS %*	*Standard Deviation*
AS units only	37.08	19.55	38.80	17.36	65.04	16.62
AS qualification only	50.77	16.46	48.23	18.12	63.70	15.08
AS qualification + A2 units	55.60	12.98	52.24	14.96	68.67	9.98
AS and A level qualifications	74.22	11.56	70.64	11.39	75.43	11.94
Not AS but A level	79.34	11.4	74.23	12.35	79.60	11.55

Table 5: Average UMS percentage in A2 units, by type of qualification obtained

Units/ Qualifications	*Biology*		*Psychology*		*English Literature*	
	Average UMS %	*Standard Deviation*	*Average UMS %*	*Standard Deviation*	*Average UMS %*	*Standard Deviation*
AS qualification + A2 units	33.19	16.92	39.42	17.51	55.36	21.38
AS and A level qualifications	67.17	15.90	64.41	15.63	73.12	13.88
Not AS but A level	72.93	15.93	66.26	15.88	77.00	13.74

In the AS units (Table 4), the worst performance in Biology and Psychology was among the candidates who dropped out before certificating for an AS level. Surprisingly, in English Literature the performance of the candidates who did not certificate for the AS level, was slightly better on average than the performance of those who did. In all three subjects the best performance was among those who achieved an A level (last two rows of Table 4). The performance of those who certificated for an AS level and attempted some A2 units was somewhere in between.

Similarly, in the A2 units (Table 5), average unit performance was better among those candidates who certificated for the A level than among those who only attempted some units and did not aggregate to

4. By inter-awarding body agreement, the uniform mark grade boundaries in AS/A level qualifications are always at the following percentages of the maximum uniform mark for the unit or qualification: A – 80%; B – 70%; C – 60%; D – 50%; E – 40%. For more details on the Uniform Mark Scale see, for example, AQA (2013).

Table 6: Percentages of candidates whose performance was worse than predicted (forecast/estimated grades) in the AS level qualification, by type of qualification obtained

Units/Qualifications	*Biology*		*Psychology*		*English Literature*	
	Number of candidates	*% performing lower than forecast*	*Number of candidates*	*% performing lower than forecast*	*Number of candidates*	*% performing lower than forecast*
AS qualification only	13,370	59.68	7,544	61.97	3,705	36.90
AS and A level qualifications	16,435	31.14	7,868	36.34	8,524	23.17
Difference		**-28.54**		**-25.63**		**-13.73**

achieve the full qualification. This pattern was consistent in all three subjects.

It is worth noting that in both AS and A2 units, candidates who did not certificate for an AS level but achieved an A level had the best average performance.

Table 6, which compares the actual and the forecast AS level grade[5] for the candidates who certificated for the AS only and those who also achieved an A level, shows that the percentages of candidates performing worse than predicted were significantly lower among candidates who continued to study the subject and achieved a full A level. This table also shows that in English Literature, the percentage of candidates with an AS only who performed worse than predicted was much lower than in Biology and Psychology.

In Biology and Psychology, a comparison between the performance in the AS subject and the performance in other attempted AS subjects showed that, for over 70% of the students (75% in Biology and 71% in Psychology) who dropped it at AS level, this subject was the one in which they achieved the lowest grade. On the contrary, the comparison between the performance in AS English Literature and the performance in other attempted AS subjects pointed out that students taking English Literature might not be dropping the subject in which they are performing worst. In fact, for more than half of the students who dropped English Literature at AS level, this was not the subject where they achieved their lowest grade.

As discussed earlier, multilevel logistic regression analyses were carried out to investigate which types of students were more likely to drop out from AS to A level. Table 7 shows the results of the regression analysis for Biology, Psychology and English Literature. The statistically significant predictors (highlighted in bold in the table) are discussed in the next section.

Table 7: Multilevel logistic regression outcomes, probability of dropping out from AS to A level

Effect		*Biology*		*Psychology*		*English Literature*	
		Estimate (SE[a])	*Odds ratio*	*Estimate (SE)*	*Odds ratio*	*Estimate (SE)*	*Odds ratio*
Intercept		**-1.02 (0.45)**	0.36	**-0.83 (0.54)**	0.44	**-1.53 (0.73)**	0.22
Gender	Male [Female]	**0.15 (0.05)**	1.16	**0.19 (0.08)**	1.21	**0.21 (0.09)**	1.23
GCSE Science subject	Biology [Additional Science]	**-0.26 (0.05)**	0.77	-	-	-	-
Subject at GCSE	No [Yes]	-	-	**0.55 (0.22)**	1.74	-0.05 (0.30)	0.96
Grade in GCSE Science subject	A B or below [A*]	0.05 (0.07) **0.29 (0.09)**	1.05 1.33	- -	- -	- -	- -
Average GCSE score		**0.58 (0.06)**	1.78	**0.57 (0.07)**	1.77	**0.62 (0.07)**	1.87
Type of schools	State-maintained Independent FE college Sixth form college [Selective]	0.07 (0.15) **1.07 (0.21)** -0.33 (0.54) **0.70 (0.30)**	1.07 2.90 0.72 2.01	-0.36 (0.31) 0.48 (0.42) -0.93 (1.26) 0.68 (0.51)	0.70 1.62 0.40 1.97	0.04 (0.21) 0.37 (0.28) - 0.14 (0.53)	1.04 1.44 - 1.16
AS level grade	A B C D E [U]	**-7.74 (0.21)** **-6.61 (0.19)** **-5.81 (0.18)** **-4.95 (0.18)** **-3.63 (0.18)**	0.14 0.45 - 2.36 8.85 334.49	**-6.49 (0.25)** **-6.05 (0.23)** **-5.31 (0.21)** **-4.63 (0.20)** **-3.18 (0.19)**	0.31 0.47 - 1.98 8.36 201.68	**-7.06 (0.57)** **-5.89 (0.56)** **-5.09 (0.55)** **-4.09 (0.55)** **-2.90 (0.56)**	0.14 0.45 - 2.70 8.94 162.15
Number of AS subjects		**0.38 (0.03)**	1.46	**0.25 (0.05)**	1.28	**0.36 (0.05)**	1.43

a. Standard Error. *Notes:* Estimates that are statistically significant at the 5% level are highlighted in bold. To aid interpretation, Odds ratios for AS level grade are given relative to grade C rather than to grade U.

5. The forecast grades submitted by the centres prior to the examinations taking place were used as the measure of predicted performance. More information about estimated/forecast grades can be found in OCR (2013d).

Gender:

In all three subjects, the gender of the student was significantly associated with the probability of dropping out from AS to A level, once the other student and school characteristics were accounted for. In particular, males were more likely to drop out than females.

GCSE subject:

In Biology, the Science subject studied at GCSE (Biology versus Additional Science) was a significant predictor of dropping out from AS to A level. In particular, against the baseline of Additional Science, candidates with GCSE Biology were less likely to drop out. In Psychology, having studied the subject at GCSE was a significant predictor of dropping out from AS to A level. In particular, those candidates who did not study for a GCSE in the subject were more likely to drop out from AS level to A level in Psychology that those with the GCSE. However, having studied for a GCSE in English Literature did not display a statistically significant association with continuing to study the subject from AS to A level.

Grade in GCSE Science subject:

The performance in Science GCSE (either Biology or Additional Science) was a significant predictor of dropping out from AS to A level. In particular, against the baseline of students who achieved a grade A*, those who achieved a grade B or below were more likely to drop out. However, students who achieved a grade A were not significantly more or less likely to drop out than students who achieved a grade A*.

Average GCSE score:

For the three subjects considered in the analyses, students with higher average GCSE scores were more likely to drop from AS to A level than students with lower average scores. This could suggest that pupils with higher prior attainment will tend to require higher grades at AS level in order to consider continuing with a subject to be worthwhile. An alternative explanation of this finding could be that students with good grades at GCSE might have taken the AS subject with the aim to broaden their curriculum but they did not consider the subject one of their core A levels.

Type of school:

In Biology, against the baseline of selective schools, candidates in independent schools and candidates in sixth form colleges were more likely to drop out from AS to A level, once the other student and school characteristics were accounted for. However, candidates in state-maintained schools or in FE colleges were not significantly more or less likely to drop out than candidates in selective schools. In Psychology and English Literature, the type of school did not display a statistically significant association with continuing to study the subjects from AS to A level.

Number of AS subjects:

In Biology, Psychology and English Literature, the number of AS subjects attempted by a student was a significant predictor of dropping out the subject from AS to A level. In particular, the higher the number of AS subjects, the higher the probability of dropping out.

AS level grade:

As expected, the performance at AS level was a significant predictor of dropping out from AS to A level. In particular, the lower the AS grade, the higher the probability of dropping out. This pattern was consistent in the three subjects. As an example, Figure 1 shows how the grade at AS changes the probability of dropping Biology from AS to A level for a girl in a state-maintained school, who achieved a grade A in GCSE Biology and had an average GCSE attainment of 6.5 (around average in this sample).

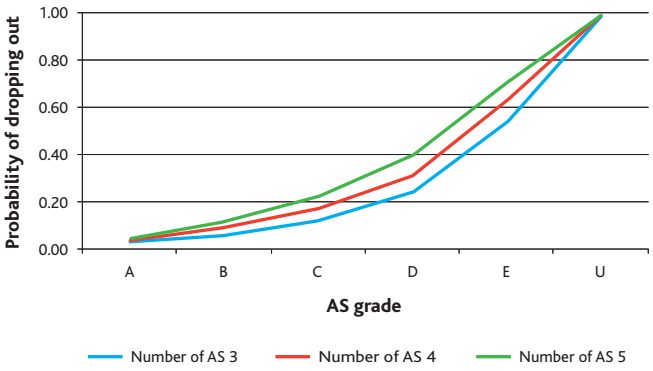


Figure 1: Effect of the AS level grade on the probability of dropping Biology from AS to A level

Summary of results and conclusions

Until now, students in England have been able to study the AS level as either a standalone qualification or as the first half of an A level. At the end of the AS year (usually Year 12), students had two options: take an AS level only and gain a recognised qualification; or continue for a second year studying the A2 units and go for the full A level.

The main aim of this research was to gain a better understanding of the numbers and types of students who decide not to continue their studies once they had started either an AS or an A level qualification. The focus was on typical AS/A level students who were at the end of KS5 (Year 13) in 2012/13 and had taken at least one AS/A level unit in the following three subjects: Biology, Psychology and English Literature.

Regarding the numbers of candidates dropping out and their performance in the AS/A2 units attempted, the analyses carried out in this article showed that:

- In all three subjects, the majority of the candidates who sat at least one unit certificated in both AS and A level. The percentage of students with both qualifications was highest in English Literature and lowest in Psychology.
- There were reasonably large percentages of candidates who had enough units for AS certification but decided not to aggregate. In most cases, aggregation would have led to a grade E or below. The exception was English Literature, where candidates who did not aggregate would have achieved, on average, a grade C.
- In the AS units, the worst performance was, in general, among those candidates who dropped out before certificating for an AS level and the best performance was among those who achieved an A level.
- In the A2 units, average unit performance was better among those candidates who certificated than among those who did not aggregate to achieve the full A level qualification.

- The percentages of candidates achieving a worse AS level grade than predicted were significantly lower among candidates who continued to study the subject and achieved a full A level than among those who dropped the subject at AS.
- In Biology and Psychology students who dropped the subject at AS level might have done so because the subject was the one in which they achieved their lowest grades. For example, for almost 75% and 71% of the students who dropped Biology and Psychology respectively, the subject was the one in which they performed the worst at AS level. However, for more than half of the students who dropped English Literature at AS level, this was not the subject where they achieved their lowest grade.

Regarding the types of students who were more likely to drop out from AS to A level, the analyses carried out in this article showed that:

- In all three subjects, boys were more likely to drop out from AS to A level than girls, once student and schools characteristics were accounted for.
- In Biology, the Science subject studied at GCSE was a significant predictor of dropping out from AS to A level. In particular, candidates who had studied a GCSE in Biology were less likely to drop out than those who had studied the GCSE in Additional Science. Furthermore, the lower the GCSE grade, the higher the probability of dropping out.
- In Psychology, the candidates who had not studied for a GCSE in the subject were more likely to drop out at AS level than those with the GCSE.
- There was no association between the type of school where the AS/A level was being studied and the likelihood of dropping out in Psychology and English Literature. However, in Biology, candidates in independent schools and in sixth form colleges were more likely to drop out from AS to A level than candidates in selective schools.
- As expected, performance at AS level was a significant predictor of dropping out from AS to A level in all three subjects. In particular, the lower the grade at AS, the higher the probability of dropping out.

  Similarly, the number of AS subjects attempted by the student was a significant predictor of dropping the subjects investigated in this study (Biology, Psychology and English Literature). In particular, the higher the number of AS subjects attempted, the higher the probability of dropping out.

In conclusion, and supporting previous research (e.g., Pinot de Moira, 2002), the results presented in this article suggest that students who dropped subjects from AS level to A level usually had a worse result for the AS part of the examination than students who continue to achieve the full A level qualification.

However, the outcomes of this work showed that an influential reason to continue to study a subject to the full A level could be the students' early interest in it (e.g., at GCSE). This research has shown, in fact, that having studied the subject at GCSE increased the likelihood of studying for a full A level rather than for just an AS level only.

Finally, it should be noted that for some Higher Education courses A level qualifications in certain subjects are required (e.g., A level Chemistry is usually a requirement to study Medicine; A level Mathematics is a requirement for Mathematics, Engineering and Physics degrees). Therefore, some subjects might be less likely to be dropped than others independently, for example, of the students' performance or enjoyment.

References

AQA (2013). *Guide to the Uniform Mark Scale (UMS)*. Manchester: AQA Education.

BBC News (2013, January 23). A-level plans challenged by school and university heads. *BBC News – Education & Family*. Retrieved from http://www.bbc.co.uk/news/education-21156370

Department for Education (2010). *The Importance of Teaching: The Schools White Paper 2010*. London: DfE. Retrieved from https://www.gov.uk/government/publications/the-importance-of-teaching-the-schools-white-paper-2010

Goldstein, H. (2011). *Multilevel Statistical Models (4th edition)*. Chichester: John Wiley & Sons.

Gove, M. (2012). *Reform of GCE A levels*. Letter to Stacey, G. [30th March 2012]. Retrieved from: http://www2.ofqual.gov.uk/files/2012-03-31-michael-gove-letter-to-glenys-stacey.pdf

Gove, M. (2013). *Reform of GCE A levels*. Letter to Stacey, G. [22nd January 2013]. Retrieved from: http://www.ofqual.gov.uk/files/24-01-2013-ofqual-letter-reform-of-gcse-a-levels.pdf

Hosmer, D.W. & Lemeshow, S. (2000). *Applied Logistic Regression*. Chichester: Wiley.

Ofqual (2014). *An update on the reforms being made to AS qualifications and A levels*. Coventry: The Office of Qualifications and Examinations Regulation.

OCR (2013a). *OCR AS/A Level GCE Biology H021/H421. Specification Version 4*. Cambridge: Oxford, Cambridge and RSA.

OCR (2013b). *OCR AS/A Level GCE English Literature H071/H471. Specification Version 4*. Cambridge: Oxford, Cambridge and RSA.

OCR (2013c). *OCR AS/A Level GCE Psychology H168/H568*. Cambridge: Oxford, Cambridge and RSA.

OCR (2013d). *Admin Guide and Entry codes: 14–19 qualifications 2013/14*. Cambridge: Oxford, Cambridge and RSA.

Pinot de Moira, A. (2002). *Preliminary Analysis of the Summer 2002 A Level Results (Internal AQA paper RC/188)*. Guildford: AQA.

Sharp, C. (1996). *Review of Qualifications for 16–19 Year Olds: completion of A level and GNVQ Courses: a Literature Review*. Slough: NFER.

Snijders, T. & Bosker, R. (1999). *Multilevel Analysis. An introduction to basic and advance multilevel modeling*. London: SAGE Publications Ltd.

Stewart, W. (2013, January 25). Clash of qualifications will result in 'big mess'. *TES Magazine*. Retrieved from https://www.tes.co.uk/article.aspx?storycode=6317702

Sutch, T. (2013). Progression from GCSE to AS and A level, 2010. *Statistics Report Series No. 69*. Cambridge: Cambridge Assessment.

Sutch, T. (2014). Uptake of GCE AS level subjects 2007–2013. *Statistics Report Series No. 75*. Cambridge: Cambridge Assessment.

Vidal Rodeiro, C.L. (2007). *A level subject choice in England: patterns of uptake and factors affecting subject preferences*. Internal Report. Cambridge: Cambridge Assessment.

Watson, L. (2013, February 2). Britain's top universities attack Government plan to hive off AS levels as a standalone qualification. *Daily Mail Online*. Retrieved from http://www.dailymail.co.uk/home/sitemaparchive/day_20130202.html

The moderation of coursework and controlled assessment: A summary

Tim Gill Research Division

Introduction

To ensure consistency and accuracy of marking, awarding bodies carry out moderation of GCSE and A level internally assessed work (e.g., coursework or controlled assessment). Training and instructions are provided by the awarding body to the internal assessors in each centre, including training in task-setting, marking and internal standardisation. Internal standardisation is necessary to ensure the standard is the same across all assessors within a centre.

Awarding bodies are required to modify centres' marks where necessary to bring judgements into line with the required standard. Samples are taken of (internally standardised) candidates' work, across all units and adequately covering the range of attainment within a centre. A moderator re-marks the sampled work, and if there is a difference between the centre's and moderator's marks that is larger than a certain amount (known as the tolerance level) then marks should be adjusted. Should it be necessary to adjust a centre's marks then the magnitude of the adjustments is determined by a regression analysis, based on the relationship between the marks given by the centre and those of the moderator in the sample.

This article summarises the processes undertaken by the Oxford, Cambridge and RSA (OCR) exam board to moderate and, if necessary, adjust the marks of centre-marked coursework and controlled assessments. Some brief data analysis is also presented to give an idea of the extent of moderation and how much difference it makes to candidates' marks.

Moderation and scaling processes

Broad guidelines for the moderation process are set out in the Ofqual Code of Practice document (Ofqual, 2011). More detailed principles and practices were drawn up by the exam boards, as described in an OCR document which provided guidance to centres (OCR, 2010). However, the processes described here refer to those undertaken by OCR only. Other boards may have different processes, so long as they comply with the Code of Practice and the board agreement.

Sampling and moderation

The Ofqual guidelines for sampling student work are quite broad, only requiring that exam boards request samples of work from centres which adequately represent the range of attainment within the centre, requesting additional samples if necessary. They do not specify how this should be done. The OCR procedures are much more detailed, as follows; for each centre taking a coursework unit a sample of (internally marked) scripts (chosen by OCR) are sent to the moderator ('Stage 1'). This sample is drawn from across the range of marks in the centre, and includes the lowest and highest centre marks. A first sample is moderated and if there are no differences above tolerance then no more moderation is necessary and the centre's marks are accepted. However, if one or more differences exceed tolerance then a further sample is moderated ('Stage 2'). If, after this second moderation, the pattern of changes suggested by the moderator is relatively consistent (i.e., it retains the rank order of candidates) then the centre's marks are scaled (see later description). If they are not consistent then it is possible to take a third sample for moderation ('Stage 3'). If after this a valid scaling is still not possible then further options include the moderator re-assessing all candidates in the centre and applying the moderated marks, or the centre re-assessing all work and a new sample being taken.

The size of the sample(s) described above depends on the number of candidates in the centre taking the coursework unit, as shown in Table 1.

Table 1: Sample sizes for different centre sizes

No. of candidates in centre	*Stage 1 (sub) sample*	*Stage 2 (full) sample*	*Stage 3 sample*
1–5	All	All	All
6–10	5	All	All
11–15	6	10	All
16–100	6	10	15
101–200	6	15	20
201+	6	20	25

Scaling

Following the moderation, the scaling adjustments that will be applied are determined through the application of a regression algorithm. The use of regression to determine adjustments is not required by the Ofqual guidelines, and in the inter-board agreement it is only given as an example of how 'automatic' adjustments could be applied. The purpose of the regression algorithm is to determine whether to adjust a centre's marks and if so, by how much. These adjustments will be applied to all candidates in the centre, not just the sample. Only centres where the result of the moderation of the sample was at least one script outside of tolerance go forward to the regression algorithm. Even then it is not certain that it will be necessary to adjust the marks in the centre. If the adjustments suggested by the algorithm are within the tolerance for the unit then the centre marks are accepted. That is, if the adjustment that would be performed would only alter marks by an amount less than or equal to tolerance, then the original marks are close enough to be accepted.

In order to decide how much to adjust a centre's marks by, a regression equation is used to model the relationship between centre and

Table 2: Example application of moderation and scaling procedure

Centre mark (X)	**36**	**36**	**34**	**33**	**31**	**29**	**27**	**24**	**21**	**18**	**16**	**16**	**11**	**7**	**6**
Moderator mark (Y)	34	32	32	31	30	29	28	23	22	21	17	16	14	8	8
Mod – Centre (Y-X)	-2	-4	-2	-2	-1	0	1	-1	1	3	1	0	3	1	2
Predicted mark (Ŷ)	33.9	33.9	32.2	31.3	29.7	28.0	26.3	23.8	21.3	18.7	17.1	17.1	12.9	9.5	8.7
Regression mark	34	34	32	31	30	28	26	24	21	19	17	17	13	10	9
Regression – Centre	-2	-2	-2	-2	-1	-1	-1	0	0	1	1	1	2	3	3
Final mark	**34**	**34**	**32**	**31**	**30**	**28**	**26**	**24**	**21**	**19**	**17**	**17**	**13**	**10**	**9**

moderator mark[1]. The form of this equation used by the algorithm is as follows:

$$Y = aX + b$$

Where Y is the moderator mark, X is the centre mark and 'a' and 'b' are the regression parameters. For each centre mark (X) in the sample a predicted adjusted mark (Ŷ, also known as the 'regression mark') is generated from this equation. The 'a' and 'b' parameters are set so as to minimise the average of the squared difference between each moderator mark and predicted mark in the sample.

The magnitude of the adjustment (if it is deemed necessary) is the difference between the centre's mark and the regression mark. Often, the regression mark is not a whole number, in which case it is rounded up or down. Take the following example, for a unit of maximum mark of 40 marks, with a tolerance of 2 marks and the following regression equation:

Moderator mark = 0.84*Centre mark + 3.62.

Table 2 presents the centre and moderator marks for the sample and follows through the procedure to get the final marks applied to these marks.

The regression equation generates a predicted mark as shown in Table 2 (Ŷ, to 1 decimal point). This is rounded up or down to generate the regression mark. This becomes the final mark if the algorithm determines that an adjustment to the centre's marks is necessary. A final check is made of whether any of the adjustments are outside of tolerance. If none are, then the centre marks are accepted. In the example in Table 2 there were two candidates with proposed adjustments of 3 marks, greater than the tolerance of 2 marks (highlighted by the red squares), so the decision would be to adjust the centre's marks.

This example is displayed graphically in Figure 1.

The crosses represent the centre and moderator mark for each candidate. The blue line is the regression line, indicating the proposed changes to centre marks. This is not straight because of the rounding up or down that is necessary to enable the mark adjustments to be whole numbers. Note that the regression line tapers at the bottom of the mark range so that candidates with a mark of zero have their mark unadjusted.

Finally, the straight lines are bands for the level of tolerance for this unit. These bands are two marks either side of the 'identity' line (not shown, but where the centre and moderator marks are equal). This figure shows that in the sample there were three scripts with marks outside of tolerance (outside of the bands). At the bottom of the mark range there were two scripts for which the regression line suggests an adjustment that would be greater than tolerance. This means that this centre's marks would need to be adjusted.

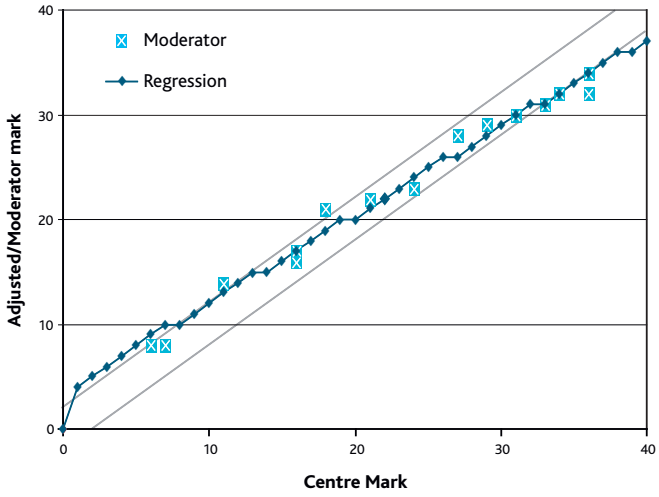


Figure 1: Plot of centre and moderator marks with regression marks, and tolerance bands

Criteria for automatic scaling

Once the algorithm has determined that adjustments are necessary, these are applied to all candidates in a centre automatically, as long as some specific criteria are met. These criteria are not required by Ofqual regulations but were generated by OCR to ensure some checks are carried out on the scaling undertaken. The aim of the criteria is to flag up any scaling decisions that are particularly out of the ordinary in some way (e.g., large adjustments to marks), or might be unfair to some candidates. If at least one of these criteria is not met, the centre is flagged up so that OCR Operations staff can look in more detail at the proposed scaling decision and decide whether or not it is valid. The criteria are:

1. No 'unusual marks' in the sample. Unusual marks are those where the difference between the regressed mark and moderator mark is larger than 10 per cent of the maximum mark.
2. The average of the squared difference between the moderator marks and the regression marks is less than or equal to 3.5. This is so that centres where the adjustments to candidates are very different to those suggested by the moderator are not included automatically.

1. For marks between 10% and 100% of the maximum mark a simple linear regression is used. For the bottom 10% the regression line is a curve so that the centre and moderator marks converge at zero.

3. No large differences between centre and moderator marks at the extremes of the sample. A large difference at the extremes might mean that excluding this candidate would have a big impact on the scaling decision and adjustments.
4. More than one mark outside of tolerance. This is because if only one mark was greater than tolerance then excluding this candidate (which is an option open to Operations staff when reviewing the recommended mark adjustments) would change the scaling decision from adjusting to not adjusting.
5. The average absolute adjustment applied to all candidates in the centre is not greater than 15 per cent of the maximum raw mark. This is to ensure that any particularly large adjustments are flagged up.
6. Correlation between centre and moderator marks is at least 0.75. A correlation lower than this would suggest a valid scaling would be difficult.

If it is decided that the proposed scaling is not valid then there are two options available. First, where there are unusual marks in the sample, it is possible to exclude these candidates and re-run the regression to see what the impact is on the proposed adjustments. If a candidate is having a detrimental effect on the adjustments for all other candidates then it might be justified to exclude them. However, candidates should not normally be excluded if it would change the scaling decision from applying to not applying adjustments.

If it is still not possible to create a valid scaling outcome using the regression algorithm then the procedures allow for manual scaling. This means manual adjustments are made at each mark point without recourse to the regression algorithm.

Once the scaling has been determined it is applied to all candidates in the centre, not just those in the sample. This is communicated to the centre in the form of a Banding Report, showing the scaling that needs to be applied to different bands of centre marks. The report covers the whole of the mark range, whether or not the centre has any candidates with a mark in a particular band. An example is shown in Table 3.

Table 3: Example Banding Report

Marks From–To	*Scaling Factor*
34–40	-2
28–33	-1
21–27	0
14–20	1
8–13	2
1–7	3

Data analysis

This section explores some background data in relation to moderation of coursework by OCR. The data comes from the June 2012 session.

Extent of moderation

Several of the qualifications offered by OCR involve some components that require moderation. Table 4 presents the number of components in each qualification that were moderated in the June 2012 session.

Table 4: Number of moderated components by qualification, June 2012

Qualification	*Components*
A level	126
GCSE	95
Principal learning (Level 1, 2 and 3)	74
Entry level certificate	18
Other	6
All	**319**

Table 5: Summary of moderated components, June 2012

Moderated components	319
Moderated centres	35,011
Regressed centres (%)	28.0
Scaled centres (%)	22.5
Candidates in scaled centres	188,091
Scaled candidates	167,763

Table 5 presents the total number of components, centres and candidates that were affected by scaling in June 2012. It also presents the percentage of centres taking units subject to moderation whose marks were scaled.

Thus, there were 319 components that were moderated and 35,011 centres subject to moderation. Of these centres, 28% were found to have at least one difference between centre and moderator mark that was larger than the allowed tolerance, meaning the marks went through the regression algorithm. However, only 22.5% of moderated centres actually had scaling applied. The reason for this difference is that some of the regressed centres had all the 'regressed' marks inside of tolerance (see earlier description). The 'candidates in scaled centres' figure in the table includes candidates whose mark did not in fact change, because their scaling adjustment was 0 marks. The 'scaled candidates' figure only includes students whose marks were adjusted.

An analysis was also undertaken of the percentage of centres in each component that were regressed. The results showed that there were 35 components where no centres were regressed (i.e., all centres marks accepted) and 18 components where all centres were regressed. Most of these components were taken by a very low number of centres (fewer than 10), but there was one component with 89 centres, none of which were regressed. There was also a component with just one centre regressed out of 191 (0.5%). Excluding components where all centres were regressed (each of which consisted of fewer than five centres), the highest percentage of centres regressed was 80.3 (196 out of 244). Figure 2 presents the distribution of the percentage regressed, for components with at least 50 centres.

In terms of the percentage of centres within each component that were actually scaled, there were 42 components where none of the centres were scaled, one of which had 191 centres and one 89 centres. Otherwise the numbers of centres for these components were generally very low. There were 13 components where all centres were scaled, but these were all components with very few centres. Of the components with more than 50 centres, the highest percentage of scaled centres was 76.7% (69 out of 90). The lowest percentage scaled was 1.2% (4 out of 343).

Figure 3 presents the distribution of the percentage scaled, for components with 50 or more centres.

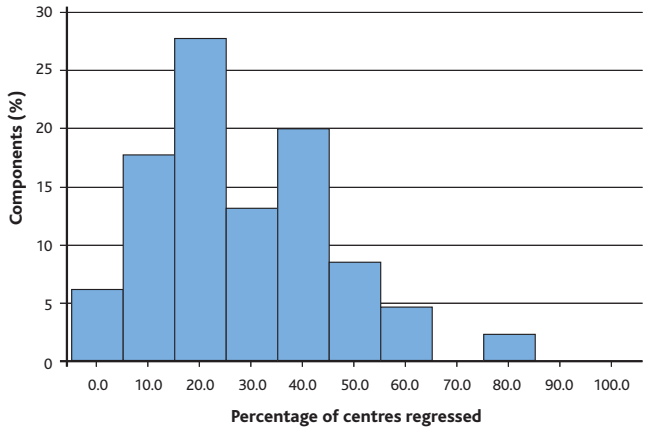


Figure 2: Distribution of the percentage of centres regressed (components with 50 or more centres)

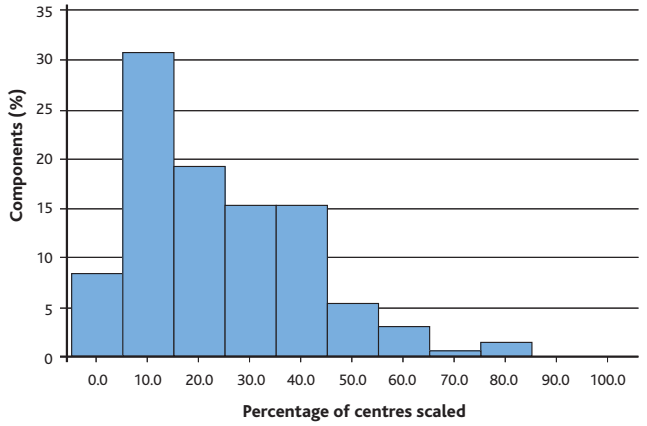
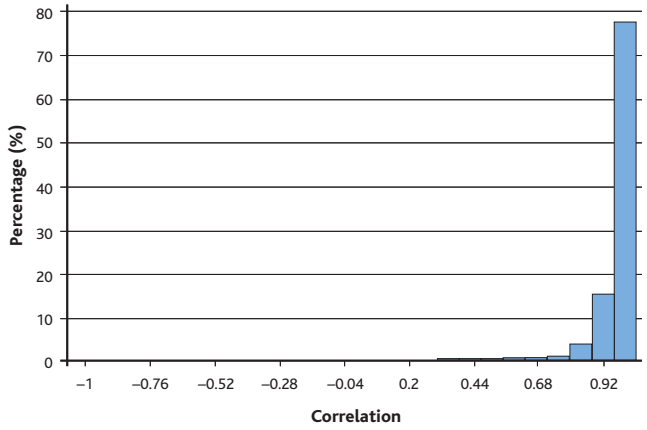


Figure 3: Distribution of the percentage of centres scaled (components with 50 or more centres)



Correlation	*Centres*	*Percentage of centres*
<0.75	135	2.0
0.75–0.80	60	0.9
0.80–0.85	119	1.8
0.85–0.90	252	3.8
0.90–0.95	656	9.8
>0.95	5,459	81.7
All	6,681	100.0

Figure 4: Distribution of correlation coefficients in scaled centres

Correlation between centre and moderator marks

One way of assessing the level of agreement (in terms of rank order of candidates) between centre and moderator marks is through a correlation coefficient. This was calculated for each (scaled) centre in each component in the June 2012 data. These correlations used the marks for the sampled scripts only, as these were the only scripts with a moderator mark. Figure 4 presents the distribution of correlation coefficients.

Almost 82% of centres had a correlation of greater than 0.95 and 91.5% had a correlation of greater than 0.90. Thus, in terms of the rank order of candidates within a centre, there is usually a lot of agreement between centre and moderator mark even in the centres which were scaled. However, this doesn't necessarily mean that there is a high level of agreement over the marks. It may be that the centre marks tend to be consistently higher than moderator marks. This explains why a substantial percentage of centres are scaled, even though correlations tend to be very high.

Adjustments to marks

Another important aspect of the scaling process is how large the adjustments to candidates' marks are. Table 6 summarises the changes to candidates' marks as a result of scaling, (a negative figure means a reduction in the mark given to the candidate). This includes all candidates in centres that were scaled, not just those in the sample.

Overall, adjustments were much more likely to be negative than positive, with just 5.8% of adjustments greater than 0, compared with 83.4% less than 0. The remaining 10.8% of candidates had no adjustments to their marks, despite being in centres where some adjustments were necessary.

Table 6: Summary of adjustments made to candidates' marks

N	*Mean*	*SD*	*Min*	*Max*
188,091	-3.9	3.9	-60	40

This analysis was repeated for the adjustment as a percentage of the maximum mark for the component. Figure 5 presents the distribution of adjustments. The mean adjustment was -6.7% with a minimum adjustment of -65% and a maximum of 58.3%.

A further analysis was undertaken of the mean adjustment to marks for each individual component. There was a fairly wide range of adjustments, with the biggest negative adjustment on average for a component being -21.6 marks (although this component was only taken

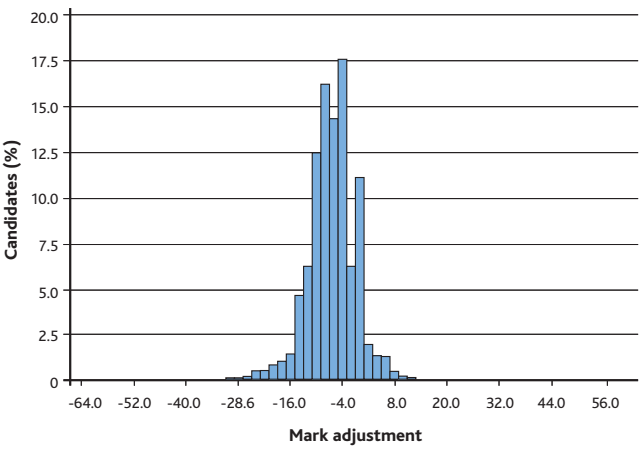


Figure 5: Distribution of mark adjustments (percentage of maximum mark)

by 19 candidates), whilst the most positive was +8.2 marks. The biggest mean adjustment for a component with more than 100 candidates was -10.8 marks. However, this component had a maximum mark of 120, so the average adjustment was less than 10%.

Figure 6 presents the distribution of mean adjustments for each component in terms of percentage of maximum mark (restricting to components with more than 100 candidates). The largest negative adjustment was -18.7% (-9.3 marks for a component with a maximum mark of 50). The largest positive adjustment was 8.2% (+8.2 marks for a component with a maximum mark of 100).

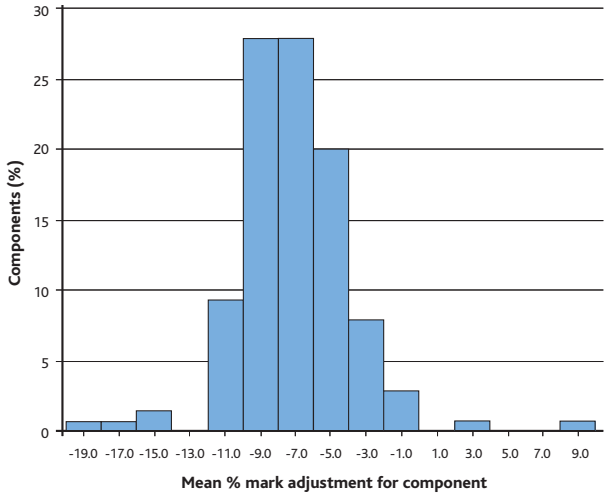


Figure 6: Distribution of mean adjustment to mark (as a percentage of maximum mark) by component

Extent of automatic scaling

As previously noted, in order for the scaling to proceed automatically without being checked, a number of criteria need to be met. For the June 2012 session the number of centres where at least one of the criteria was failed and the scaling outcome was checked was 2,102 (31.3% of all centres scaled). Table 7 presents the number of centres failing each criterion. The criteria are not mutually exclusive, so it is possible for centres to fail more than one.

Table 7: Frequency of centres where criteria for automatic processing not met

Criterion	*Count*	*Percent*
1	108	1.61
2	980	14.58
3	403	5.99
4	482	7.17
5	676	10.06
6	139	2.07

Of the centres that were checked, 14.6% (306 centres) had their scaling adjusted manually (either by excluding candidate(s) from the regression and re-running or by deciding on the scaling to be applied at each mark point without recourse to the regression algorithm). This is 4.6% of all centres that were scaled.

Table 8 shows the frequency of the centres where a given number of criteria were not met. For instance, the first row shows that all the six criteria were met in 68.73% of centres. Around 23% of centres failed to meet just one criterion and relatively few centres (7.75%) failed on two or more of the criteria. The final row in the table indicates that 2,102 centres failed at least one criterion.

Table 8: Frequency of all criteria not met, in centres that were scaled

Count of criteria	*Count*	*Percent*	*Cumulative count*	*Cumulative Percent*
0	4,621	68.73	4,621	68.73
1	1,581	23.52	6,202	92.25
2	387	5.76	6,589	98.01
3	106	1.58	6,695	99.58
4	25	0.37	6,720	99.96
5	3	0.04	6,723	100.00
At least 1	2,102	31.27	-	-

Discussion

This article has outlined the purpose and processes involved in the moderation of coursework and controlled assessment at OCR. It has also demonstrated the extent of moderation undertaken, both in terms of the percentage of centres moderated and the levels of adjustments implemented.

It is worth noting that moderation is not meant to be the same as re-marking of work. It would not be possible for all the work in a centre to be re-marked because of the number of candidates taking these units. Instead, as described earlier, moderators re-mark a sample of the work, and use the relationship between moderator mark and centre mark in the sample to estimate what adjustments should be made to candidates' work in the whole centre. This means that some candidates whose work had been moderated will end up with a mark that is different from the mark they 'should' have received (as given by the moderator). However, the principle here is to be as fair as possible to all candidates in the centre, including those whose work has not been moderated. As we don't know the actual mark these candidates should have received, the best estimate is that generated by the relationship between moderator and centre mark (as long as that relationship is reasonably consistent across the mark range).

This article has shown that most centre marking is of the required standard: less than one quarter (22.5%) of centres taking moderated components needed to have their marks adjusted. Furthermore, when adjustments were necessary, these tended to be small (although there were some exceptions) and the correlations between centre and moderator mark (within a centre) were mostly very high. This suggests that the guidelines and training given to assessors within centres by OCR (in terms of marking and internal standardisation) are generally clear and understandable. We have also shown that only around 1 in 7 (14.6%) scaling decisions that were flagged as requiring checking were subsequently changed. This suggests that, on the whole, the regression algorithm works well in generating fair adjustments to candidates' marks.

However, it is also worth noting that it was much more common for centres to be generous than severe in their marking, in comparison to the moderator mark. This is perhaps not surprising, as teachers want their students to do as well as possible in their qualifications.

Finally, two further points about how OCR ensures that moderation is as fair and accurate as possible are worth mentioning. First, Ofqual regulations require that moderators must be trained and undertake

standardisation and have their moderation standards checked by a senior moderator. Those judged to be unsatisfactory will no longer be allowed to undertake moderation and candidates' work in centres that they moderated will need to be re-moderated. Secondly, if a centre has its candidates' work scaled and is unhappy with the adjustments made, they can request a review of the moderation (for a fee). If it is determined that the original moderation is not acceptable then a revised moderation is implemented instead.

References

Ofqual (2011), *GCSE, GCE, Principal Learning and Project Code of Practice: May 2011*. Coventry: The Office of Qualifications and Examinations Regulation. Retrieved from http://ofqual.gov.uk/documents/gcse-gce-principal-learning-and-project-code-of-practice/

OCR (2010). *Moderation of GCE, GCSE and FSMQ centre-assessed units/components: Common principles and practices*. Cambridge: Oxford, Cambridge and RSA.

Reflections on a framework for validation – Five years on

Stuart Shaw Cambridge International Examinations and **Victoria Crisp** Research Division

Abstract

In essence, validation is simple. The basic questions which underlie any validation exercise are: what is being claimed about the test, and are the claims warranted (given all of the evidence). What could be more straightforward? Unfortunately, despite a century of theorising validity, it is still quite unclear exactly how much and what kind of evidence or analysis is required in order to establish a claim to validity. Despite Kane's attempts to simplify validation by developing a methodology to support validation practice, one which is grounded in argumentation (e.g., Kane, 1992), and the "simple, accessible direction for practitioners" (Goldstein & Behuniak, 2011, p.36) provided by the *Standards* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA, APA, & NCME], 2014), good validation studies still prove surprisingly challenging to implement.

In response, a framework for evidencing assessment validity in large-scale, high-stakes examinations and a set of methods for gathering validity evidence was developed in 2008/2009. The framework includes a number of validation questions to be answered by the collection of appropriate evidence and by related analyses. Both framework and methods were piloted and refined. Systematic implementation of the validation framework followed which employs two parallel validation strategies:

1. an experimental validation strategy which entails full post-hoc validation studies undertaken solely by research staff
2. an operational validation strategy which entails the gathering and synthesis of validation evidence currently generated routinely within operational processes.

Five years on, a number of issues have emerged which prompted a review of the validation framework and several conceptual and textual changes to the language of the framework. These changes strengthen the theoretical structure underpinning the framework.

This paper presents the revised framework, and reflects on the original scope of the framework and how this has changed. We also consider the suitability and meaningfulness of the language employed by the framework.

Validation: a task too far?

Samuel Messick's extended account of validity and validation came to dominate the educational and psychological measurement and assessment landscape of the 1980s and 1990s. Instigated by Loevinger (1957), developed and articulated by Messick (1989), and endorsed through the support of significant allies including Robert Guion, Mary Tenopyr and Harold Gulliksen, the essence of validity came to be understood as being fundamentally a unitary concept. Messick's landmark treatise on validity published in the textbook *Educational Measurement* (Messick, 1989) represented the culmination and enunciation of a paradigm shift towards a unified view of validity as articulated in the description of modern construct validity. Measurement was to assume centre stage and came to be the foundation for all construct validity. Since that time, mainstream scholars have consistently affirmed the 'consensus' concerning the nature of validity (e.g., Shepard, 1993; Moss, 1995; Kane, 2001; Downing, 2003; Sireci, 2009) described in the maxim: all validity is construct validity. If validity pivots upon score meaning then by extension construct validation, that is, scientific inquiry into score meaning, is to be understood as the foundation for all validation inquiry. Hence, "... all validation is construct validation." (Cronbach, 1984, p.126).

Tests were to be evaluated holistically, on the basis of a scientific evaluation into score meaning. This approach was to have profound implications for all validation effort. Messick (1998, pp.70–71) seemed to imply that every kind of validation evidence is not only *relevant* but also *necessary* for every validation. Construct validation was to entail scientific theory-testing premised on multiple evidential sources. If the scope of modern validity theory was to be enlarged in an attempt to embrace a full evaluative treatment of consequences (as many, though not all, leading theorists of the day argued and continue to argue) then validation would require monumental effort especially if it was to include an exploration of unintended consequences.

The argument-based approach to validation – as championed by Kane (e.g., 1992, 2001, 2004, 2006, 2013), was an attempt to simplify both validity theory and validation practice. Recognising the difficulties in translating construct validity theory into construct validation practice, Kane rejects the idea that all kinds of evidence are required for every

validation exercise (thereby running counter to the ethos of the construct validity thesis). He introduced the idea that test score interpretation is defined as an interpretive argument[1], which serves to identify assessment inferences and their sources of evidence. The interpretive argument provides a generic version of the proposed interpretation and use of scores, which can be applied to some population of interest. Kane asserts that the structure of the interpretive argument and the inferences and assumptions it necessarily entails, depend on the type of interpretation to be validated. Different interpretive arguments necessarily entail different patterns of inference. More ambitious theory-based interpretations require more evidence than less ambitious ones (Kane, 2009, 2013). Accordingly, certain kinds of evidence are irrelevant to validation relating to certain kinds of proposed interpretations and score uses.

Part of the persuasive power of the interpretive argument is the guidance it allegedly provides would-be practitioners. Although Kane's argument-based approach is widely regarded as a positive development, there have been few examples of its implementation. Even fewer examples of validity arguments for large-scale educational assessments are available to the research community (Goldstein & Behuniak, 2011). Where examples are published, they tend to lack a strong evaluative dimension (Haertel & Lorie, 2004; Kane, 2006), fall short of providing a compelling argument (Sireci, 2009, p.33), and fail to demonstrate how a test is constructed to represent a construct independent of test use (Sijtsma, 2010, p.782).

Summarising the period over the last thirty years, the modern construct validity 'consensus' appears to have engendered a legacy of unresolved tensions between those for whom the practice of validation is "a lengthy, even endless process" (Cronbach, 1989, p.151) and those with a responsibility for test development to provide sufficient, general validity evidence (of the instrumental value) attesting to the quality of their measurement procedures.

Notwithstanding the now, near universal acceptance of the modern unified conception of validity there remains a lack of coherence between theory and practice (e.g., Jonson & Plake, 1998; Hogan & Agnello, 2004; Cizek et al., 2008; Shaw, Crisp & Johnson 2012), or, as Messick put it, a "persistent disjunction between validity conception and validation practice" (Messick, 1988, p.34). Early in the twenty-first century the practice of validation still remains somewhat "impoverished" according to Brennan (2006, p.8) though there are pockets of good practice (e.g., Sireci, et al., 2006; Shaw & Weir, 2007; Chapelle, Enright & Jamieson, 2008; Khalifa & Weir, 2009; Sireci, 2012).

Kane's (2006) theorisation of the interpretive argument for traits

In Kane's (2006) seminal chapter in *Educational Measurement* he set out the interpretive argument implicit in trait interpretations. The core of his visualisation for the interpretation of traits is represented in Figure 1 (based on Kane, 2006, p.33, Figure 2.2). This illustrates the basic notion of making inferences from student performance through to the domain (and traits) of interest.

So for any assessment, the students conduct the tasks given which results in evidence of their performance on those specific tasks in that

1. In 2013, Kane decided to abandon the label 'interpretive argument' in favour of interpretation/use arguments (IUAs) because the old formulation had given insufficient weight to uses. The new formulation also usefully allows a distinction to be made between interpretation and use arguments.

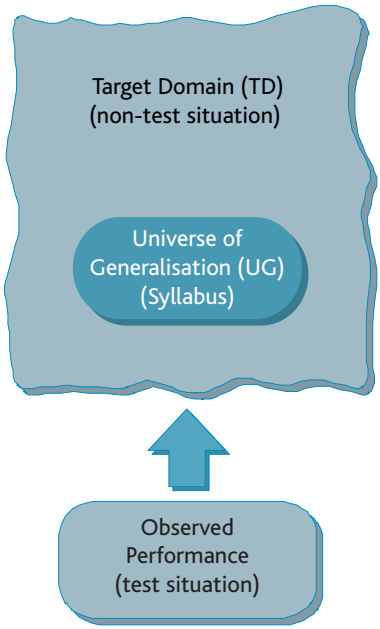


Figure 1: Interpreting traits from performance (from Kane, 2006)

testing situation. Usually, the intended uses and interpretations of the results from an assessment (or from several assessments making up a qualification) mean that stakeholders need to make inferences about competence beyond those specific tasks. In other words, stakeholders need to know that this tells us something about how much of the relevant trait(s) each student has, both:

a. specifically in relation to the range of tasks that the assessment might reasonably have encompassed (based on the content and skills set out in the syllabus) and which scores are intended to represent – this is termed by Kane the 'Universe of Generalisation' (UG) and
b. more broadly to the domain of *any* possible tasks relating to the trait – this is termed the 'Target Domain' (TD) (of which the more limited Universe of Generalisation is a subset).

According to Kane (2006), "the Universe of Generalisation for the measure of a trait is often a small subset of the target domain and tends to be defined more precisely than the target domain" (p.34). The target domain can be thought of as the domain of interest in which the ability/abilities would be observed. The target domain goes beyond the scope of the testing situation to other tasks that could have been included in the assessments given the syllabus, and beyond to trait-relevant tasks in further study or employment contexts; in other words, a broader domain of non-assessment tasks and non-assessment contexts.

This underpinning notion of the interpretation of traits from performance and Kane's argument structure underpinned the validation framework development to be described in this article.

Proposing a validation strategy for large-scale, high-stakes international examinations

Following the development of an initial draft validation framework and set of methods for gathering validity evidence, the framework was piloted in 2008 with an International A level Geography qualification (Phase 1). This resulted in a number of revisions to the framework and proposed methods, involving streamlining the subset of methods used on the basis

of how useful they were in providing evidence to evaluate validity and on the basis of their practicality.

In 2009 the framework (shown in Figure 2) was used to build a validity argument for an International A level Physics qualification (Phase 2). The framework provided the structure for collecting evidence to support the claim for the validity of the qualification, and to identify any potential threats to validity for this qualification such that they could be addressed. The structure of the validity argument was presented as an operationally-orientated validity portfolio which comprised details of the interpretive argument, validity evidence, and an evaluation of the validity argument.

The final phase of the developmental work attempted to ascertain how best to operationalise future validation effort. Through extensive consultation with colleagues and reflection on the experiences of the first two phases, Phase 3 aimed to provide suggestions for how to move forward with a strategy for validation of assessments.

A number of alternative validation strategies, from the stance and perspective of an international awarding body, were explored. These ranged in the degree to which they would provide sound evidence of the validity of assessments, and in the amount of resourcing that would be required. Whilst an attempt was made to develop streamlined and efficient methods, it was recognised that a robust evaluation of the validity of a qualification inevitably requires significant resource. The strategy adopted provided a practical and strategic approach to validity and validation where two approaches are undertaken in parallel:

1. an experimental strategy in which researchers conduct a full post-hoc validation of one or more syllabuses each year (or as necessary)

   plus

2. an operational strategy to be gradually introduced for all syllabuses, designed to gather and synthesise validation evidence currently generated routinely within an operational and assessment context.

Following implementation of the dual strategic approach to validation, a number of issues have emerged which have triggered not only a review of the validation methods but also the nature, scope and remit of the validation framework – in particular the questions addressed by the framework and the language employed in the framework.

Structure for the argument of assessment validation

The framework involves a list of inferences to be justified as indicated by a number of linked validation questions, each of which is to be answered by the collection of relevant evidence. The validation framework invites the collection of a considerable body of information in relation to categories of evidence presented in the fourth and fifth editions of the *Standards* (AERA, APA, & NCME, 2014); yet it ultimately adopts Kane's argument-based approach (e.g., Kane, 2006) in order to structure and judge that information.

Drawing on Kane's chain of inferences, the framework incorporates an underpinning logic for constructing an 'interpretive argument' (statements of claimed inferences from assessment outcomes, and the warrants which justify the inferences) based on a core structure common to all interpretive arguments within educational measurement, for the purpose of establishing measurement quality: performance inference (*Construct representation*), scoring inference, generalisation inference, extrapolation inference. In addition, a decision-making inference is included which

Figure 2: Framework for the argument of assessment validation

Interpretive argument		*Validity argument*	*Evaluation*	
Inference	*Warrant justifying the inference*	*Validation questions*	*Evidence for validity*	*Threats to validity*
Construct representation	**Tasks elicit performances that represent the intended constructs**	1. Do the tasks elicit performances that reflect the intended constructs?		
Scoring	**Scores/grades reflect the quality of performances on the assessment tasks**	2. Are the scores/ grades dependable measures of the intended constructs?		
Generalisation	**Scores/grades reflect likely performance on all possible relevant tasks**	3. Do the tasks adequately sample the constructs that are set out as important within the syllabus?		
Extrapolation	**Scores/grades reflect likely wider performance in the domain**	4. Are the constructs sampled representative of competence in the wider subject domain?		
Decision-making	**Appropriate uses of scores/grades are clear**	5. Is guidance in place so that stakeholders know what scores/ grades mean and how the outcomes should be used?		

Evaluation of claim		
	Evidence for validity	*Threats to validity*
How appropriate are the intended interpretations and uses of test scores?		
Interpretation 1. Scores/grades provide a measure of relevant learning/achievement		
Interpretation 2. Scores/grades provide an indication of likely future success		

enables a decision to be taken about test takers on the basis of their score. These inferences make up an interpretive chain which flows[2]:

from the task to the test performance (*Construct representation*)

↓

from the test performance to the test score/grade (*Scoring*)

↓

from test score/grade to test competence (*Generalisation*)

↓

from test competence to domain competence (Extrapolation)

↓

from domain competence to trait competence (*Decision-making*)

2. See Tables 1 to 5 on pages 34–35 for definitions of terms.

For each inference, an associated warrant sets out a statement that is claimed to be true. The warrant, if appropriately supported by evidence through the validity argument, justifies the intended inference to which it relates.

The findings of validation exercises based on the framework would present 'Evidence for validity' and any potential 'Threats to validity'. Any identified threats to validity might provide advice for test development in future sessions, or might suggest recommendations for changes to an aspect of the qualification, its administration and procedures or associated documentation. The second table within the framework facilitates making conclusions about whether the intended interpretations of assessment outcomes (as set out in test claims) are appropriate given the evidence collected. For a full description of the development of the framework see Shaw, Crisp and Johnson (2012) and Shaw and Crisp (2012).

Framework revisions: issues and challenges

Implementation of the framework – designed to be used in the context of traditional written examinations (within general, academic qualifications) – revealed the emergence of a number of issues. The issues relate to the way in which the *Generalisation*, *Extrapolation* and *Decision-making* inferences are conceptualised and articulated. The *Construct representation*, and *Scoring* inferences remained unchanged in meaning and terminology as no issues had arisen in relation to these. Tables 1 and 2 set out the details of these two inferences for reference along with some brief explanation. The conceptual and linguistic revisions made to the framework in the remaining three inferences will then be described and are tabulated in Tables 3–5 (revisions are shown in red highlight in column 2).

Table 1: Construct representation inference

CONSTRUCT REPRESENTATION	
from Tasks to Test performance	
Test performance =	profile of performance on test tasks
Warrant:	Tasks elicit performances that represent the intended constructs
Validation question:	Do the tasks elicit performances that reflect the intended constructs?
Infer that:	The tasks elicit the intended test constructs.

Table 2: Scoring inference

SCORING	
from Test performance to Test score/grade	
Test score/grade =	mark total across all papers within syllabus (and related grade)
Warrant:	Scores/grades reflect the quality of performances on the assessment tasks
Validation question:	Are the scores/grades dependable measures of the intended constructs?
Infer that:	Test scores/grades represent intended constructs and quality of performance.

The Construct representation inference begins with the assessment tasks which (it is hoped) elicit performances representing the constructs of interest. Here, the validation question relates to whether the intended constructs are indeed reflected in the performances that are elicited. This is the first step in allowing stakeholders to make interpretations from performance (observed performance in the test situation) to the student's traits.

The Scoring inference relates to whether the scores or grades are a dependable measure of the intended constructs and reflect quality of performance in those constructs.

The Generalisation inference

The Generalisation inference advances the interpretive argument with a warrant that the test score/grade represents what would be obtained in the Universe of Generalisation (UG), that is, in all possible tasks that could fall within the scope of the syllabus. Generalisation depends on the "representativeness of the sample of observations and about the adequacy of the sample size for controlling sampling error." (Kane, 2006, p.34). If the test score/grade is an indication of expected performance over a domain of similar task performances all of which can be drawn from the content of the subject syllabus, then the syllabus itself constitutes the Universe of Generalisation. The syllabus is designed to reflect a view of the knowledge, understanding and skills that it is appropriate to develop in students at the level being assessed and is consistent with the current (or desired) curricular framework for the students for whom it is intended. Thus a claim relating to how well a test taker performs on a particular set of tasks on a particular occasion can be generalised to claims about expected performance on a larger domain of tasks drawn from the syllabus content (a universe of possible observations).

Table 3 shows details of the Generalisation inference in the validation framework before and after recent changes. In this inference, the student's overall competence in all tasks that could fall within the syllabus is inferred from the test score/grade. Changing from using the term *Test competence* in the original framework to *Syllabus competence* was intended to clarify the intended meaning of this term. The label 'Test competence' was considered too limiting in terms of the claims made about test taker performance and appeared to fail to convey its intended

Table 3: Generalisation inference – conceptual and linguistic changes

ORIGINAL	*REVISED*
from Test score/grade to Test competence	**from Test score/grade to Syllabus competence**
Test competence = overall competence in subject (all relevant subject tasks within scope of the syllabus)	**Syllabus competence** = overall competence in relation to all tasks that could be tested within the scope of the syllabus
Warrant: Scores/grades reflect likely performance on all possible relevant tasks	**Warrant**: Scores/grades reflect likely performance on all possible relevant tasks
Validation question: Do the tasks adequately sample the constructs that are set out as important within the syllabus?	**Validation question:** Do the tasks adequately sample the constructs that are set out as important within the syllabus?
Infer that: The scores on the tasks reflect scores on other tasks within the domain (expected scores).	**Infer that:** The scores/grades on the tasks reflect scores on other tasks within the syllabus.

meaning sufficiently clearly to evaluators new to the framework. For example, one evaluator expected Test competence to refer only to competence in the specific tasks in the specific assessment(s) which is not an unreasonable interpretation of the term. Thus, the term was adjusted to more clearly include the broader domain represented by the syllabus. The new term Syllabus competence was hoped to help with understanding, but does not represent a change to the underpinning meaning of the elements of the generalisation inference. In the revised framework, scores on test tasks reflect scores on other tasks within the syllabus (Table 3).

The Extrapolation inference

Extrapolation is central to educational and psychological assessment (Newton, 2013) and advances the interpretive argument further. The Extrapolation inference moves beyond reporting measures of observed performance in a relatively narrow domain to interpreting these more widely. The Extrapolation inference is an extrapolation to a broader domain of tasks (the Target Domain – TD) with the warrant that the universe score is what would be obtained in the TD and is used to predict future performance in some other, different context such as further study or employment.

In other words, extrapolation is an indication of likely wider performance beyond the local assessment context and suggests broader competence within and beyond the subject. The observed score can be interpreted as an indication of performance in the target setting (e.g., Higher Education Institution or workplace). Extrapolation translates performance in a local context (the test situation) into a prediction of performance in a future, non-test situation. How closely that future context relates to the knowledge and skills represented by the syllabus will affect how strong an indication of performance we can reasonably expect scores/grades to be. For example, a student's result for A level Physics is likely to be a stronger indicator of future performance on a Physics degree course, than on a Sociology degree course, and is likely to be a stronger indicator of likely future performance in a career as an Engineer than in a career as a Human Resources Consultant.

Domain competence in the original framework related only to overall competence in the subject, that is, it included competencies represented within the syllabus and going beyond this to wider competence in the subject area. However, having used this term in the framework for several years it became apparent that the meaning was not entirely understood. Validators using the framework for the first time were unsure whether it should be interpreted as the domain of the syllabus or the domain of the subject. Also, through implementation and reflection it was unclear whether this inference included just extrapolation to subject competence or extrapolation to the subject and beyond. As a result of extensive consultation and further review of the literature, it was decided that the extrapolation link should relate to making inferences from competence in the syllabus to competence in the wider subject and beyond, though it is expected, of course, that scores/grades would give a weaker indication of the latter than the former. Thus, the term *Broad competence* was chosen and the warrant, validation question, and explanation adjusted to reflect this. *Broad competence* widens the concept and relates to overall competence *within and beyond* the subject (Table 4). Accordingly, the validation question is broadened in the revised framework to include related competence beyond the subject. Enlarging the concept has implications for validation practice; because the scope of the interpretation is enhanced, "new kinds of evidence for support

Table 4: Extrapolation inference – conceptual and linguistic changes

ORIGINAL	*REVISED*
from Test competence to Domain competence	**from Syllabus competence to Broad competence**
Domain competence = overall competence in subject	**Broad competence** = overall competence within and beyond the subject
Warrant: Scores/grades reflect likely wider performance in the domain	**Warrant:** Scores/grades give an indication of likely wider performance
Validation question: Are the constructs sampled representative of competence in the wider subject domain?	**Validation question:** Do the constructs sampled give an indication of broader competence within and beyond the subject?
Infer that: The scores in tasks within syllabus domain reflect wider competencies in the subject.	**Infer that:** The scores/grades in tasks the within the scope of the syllabus give an indication of wider competencies in the subject and beyond.

(e.g., criterion-related studies or analyses of the commonalities between assessment performance and performance in the wider domain)" (Kane, 2011, p.8) are required.

The Decision-making inference

The Decision-making inference advances the interpretive argument still further by allowing decisions to be made on the basis of test scores/grades by inferring that these give an indication of preparedness for further study/work. Reflecting on the concepts inherent in the original framework, it was thought that adjustment of the Decision-making inference was needed to accommodate the broadening of the Extrapolation inference to beyond the subject area. The shift of emphasis from guidance to aiding appropriate decision-making appears to be a positive step (Table 5).

The link back to decisions has also been made clearer in the validation question: Do scores/grades give an indication of success in further study or employment such that they can be used to make appropriate decisions? Appropriate decisions can only be made if the meaning of test scores is clearly interpretable by a raft of relevant stakeholders. Clear guidance to

Table 5: Decision-making inference – conceptual and linguistic changes

ORIGINAL	*REVISED*
from Domain competence to Trait competence	**from Broad competence to Preparedness for future study/work**
Trait competence = readiness for studying the subject (or another subject) at a higher level (e.g., university study), and aptitude for work in a related field	**Preparedness for future study/work** = preparedness for further study in the subject (or another subject), and aptitude for work
Warrant: Appropriate uses of scores/grades are clear	**Warrant:** Scores/grades give an indication of likely success in further study or employment
Validation question: Is guidance in place so that stakeholders know what scores/grades mean and how the outcomes should be used?	**Validation question:** Do scores/grades give an indication of success in further study or employment such that they can be used to make appropriate decisions?
Infer that: A student's likely future success in education and employment in relevant fields.	**Infer that:** The scores/grades on the tasks give an indication of a student's future success in education and employment and can be used to make appropriate decisions.

university admissions staff, for example, will facilitate admissions and placement decisions, thus exam board guidance on score/grade meaning would still be one source of evidence to be used in answering this validation question.

In the original version of the framework the term *Trait competence* was used in this inference to refer to readiness for further study and aptitude for work. On reflection, it was thought that the notion of 'readiness' or 'preparedness' was felt to be key and not well represented by the term 'Trait competence', hence the change to 'Preparedness for future study/work'. In the Decision-making inference the notion of competence and preparedness going beyond the specific area of study is continued from the previous inference (e.g., that a good grade in one subject can provide some level of indication of preparedness for study or work in related or less related fields).

The logical structure of an interpretive argument is valuable in the context of evaluating validity as awarding bodies are effectively making a *claim* that an assessment is valid, which needs to be backed by *evidence* (derived from theory, prior research or professional experience, or from evidence gleaned specifically as part of validation operations) via a *warrant* (justifying the inference), in order to defend the claim of validity against *rebuttals* (alternative explanations, or counter claims to the intended inference). (See Toulmin's Model of Inference, 1958/2003.)

Each inference depends on a number of assumptions which require different types of backing evidence relevant to the inference. Decision-making inferences generally rely on assumptions about the appropriateness of decisions made on the basis of test scores at the individual level. The evidence relevant to the Decision-making inference may include questionnaires to stakeholders devised in order to explore how Higher Education lecturers, undergraduate students and secondary school teachers understand and use test outcomes (e.g., scores/grades).

Revised interpretive chain

The full extent of the edits made to the original framework (specifically the validation questions and warrants) is shown in Figure 3. The revised interpretive chain now flows:

from the task to the test performance (*Construct representation*)

↓

from the test performance to the test score/grade (*Scoring*)

↓

from test score/grade to syllabus competence (*Generalisation*)

↓

from syllabus competence to broad competence (*Extrapolation*)

↓

from broad competence to preparedness for future study/work (*Decision-making*)

Concluding comments

This article has described a number of revisions to an established framework designed for evidencing assessment validity in large-scale, high-stakes international examinations. The original framework has recently been subject to a review resulting in a number of conceptual and textual changes. It is believed that the changes not only strengthen the theoretical structure underpinning the framework but also ensure that the framework is more transparent in terms of the clarity of its interpretive argument.

The structure for supporting validation was designed for traditional, awarding-based written examinations. These examinations can be characterised as a 'review and award' model (Section 3 of the *Cambridge Approach*, Cambridge Assessment, 2009). Other forms of established assessments (e.g., for vocational qualifications) and the emergence of other more innovative, technologically-driven forms of assessment such as twenty-first century skills (e.g., collaborative problem-solving, creativity and decision-making) and computer-based testing will only make the process of validation more complex. Indeed, the conceptual changes and textual edits described here actually make validation more of a challenge for the validator. Nevertheless, the challenge of validation – no matter how great, should not impede its continuing execution.

Figure 3: Revised framework for the argument of assessment validation

Interpretive argument		*Validity argument*	*Evaluation*	
Inference	*Warrant justifying the inference*	*Validation questions*	*Evidence for validity*	*Threats to validity*
Construct representation	**Tasks elicit performances that represent the intended constructs**	1. Do the tasks elicit performances that reflect the intended constructs?		
Scoring	**Scores/grades reflect the quality of performances on the assessment tasks**	2. Are the scores/ grades dependable measures of the intended constructs?		
Generalisation	**Scores/grades reflect likely performance on all possible relevant tasks**	3. Do the tasks adequately sample the constructs that are set out as important within the syllabus?		
Extrapolation	**Scores/grades give an indication of likely wider performance**	4. Do the constructs sampled give an indication of broader competence within and beyond the subject?		
Decision-making	**Scores/grades give an indication of likely success in further study or employment**	5. Do scores/grades give an indication of success in further study or employment such that they can be used to make appropriate decisions?		
Evaluation of claim				
		Evidence for validity	*Threats to validity*	
How appropriate are the intended interpretations and uses of test scores?				
Interpretation 1. Scores/grades provide a measure of relevant learning/achievement				
Interpretation 2. Scores/grades provide an indication of likely future success				

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, APA, & NCME). (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Brennan, R.L. (2006). Perspectives on the evolution and future of educational measurement. In R.L. Brennan (Ed). *Educational Measurement* (4th edition) (pp.3–16). Washington, DC: American Council on Education/Praeger.

Cambridge Assessment. (2009). *The Cambridge Approach: Principles for designing, administering and evaluating assessment*. Cambridge: A Cambridge Assessment Publication.

Chapelle, C.A., Enright, M.K., & Jamieson, J.M. (2008). *Building a validity argument for the Test of English as a Foreign Language*. Abingdon: Routledge.

Cizek, G.J., Rosenberg, S.L., & Koons, H.H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, *68*(3), 397–412.

Crisp, V. & Shaw, S. (2012). Applying methods to evaluate construct validity in the context of A level assessment. *Educational Studies*, *38*(2), 209–222.

Cronbach, L.J. (1984). *Essentials of psychological testing*. New York: Harper & Row.

Cronbach, L.J. (1989). Construct validation after thirty years. In R.L. Linn (Ed.). *Intelligence: Measurement, Theory and Public Policy* (pp.147–171). Urbana: University of Illinois Press.

Downing, S.M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, *37*(9), 830–837.

Goldstein, J. & Behuniak, P. (2011). Assumptions in alternate assessment: An argument-based approach to validation. *Assessment for Effective Intervention*, *36*(3), 179–191.

Haertel, E.H. & Lorie, W.A. (2004). Validating standards based test score interpretations. *Measurement: Interdisciplinary Research and Perspectives*, *2*, 61–104.

Hogan, T.P. & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement*, *64*(4), 802–812.

Jonson, J.L. & Plake, B.S. (1998). A historical comparison of validity standards and validity practices. *Educational and Psychological Measurement*, *58*(5), 736–753.

Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535.

Kane, M.T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342.

Kane, M.T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, *2*(3), 135–170.

Kane, M.T. (2006). Validation. In R.L. Brennan (Ed). *Educational Measurement* (4th edition) (pp.17–64). Washington, DC: American Council on Education/Praeger.

Kane, M.T. (2009). Validating the interpretations and uses of test scores. In R.W. Lissitz (Ed.). *The Concept of Validity: Revisions, new directions, and applications* (pp.39–64). Charlotte, NC: Information Age Publishing, Inc.

Kane, M.T. (2011). Validating score interpretations and uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing*, *29*(1), 3–17.

Kane, M.T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Khalifa, H. & Weir, C.J. (2009). *Examining Reading: Research and practice in assessing second language reading*. Cambridge: Cambridge University Press.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635–694.

Messick, S. (1988). The once and future issues of validity: assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp.33–45). Hillsdale, NJ: Lawrence Erlbaum Associates.

Messick, S. (1989). Validity. In R. Linn (Ed.). *Educational Measurement* (3rd edition) (pp.13–100). Washington, DC: American Council on Education.

Messick, S. (1998). Alternative modes of assessment, uniform standards of validity. In M. Hakel (Ed.). Beyond multiple choice: *Evaluating alternatives to traditional testing for selection* (pp.59–74). Mahwah, NJ: Lawrence Erlbaum Associates.

Moss, P.A. (1995). Themes and variations in validity theory. *Educational Measurement: Issues and Practice*, *14*(2), 5–13.

Newton, P.E. (2013). Two kinds of argument. *Journal of Educational Measurement*, *50*(1), 105–109.

Shaw, S. & Crisp, V. (2012). An approach to validation: Developing and applying an approach for the validation of general qualifications. *Research Matters: A Cambridge Assessment Publication*, *Special Issue 3*, 1–44.

Shaw, S., Crisp, V. & Johnson, N. (2012). A framework for evidencing assessment validity in large-scale, high-stakes international examinations. *Assessment in Education: Policy, Principles & Practice*, *19*(2), 159–176.

Shaw, S. & Weir, C.J. (2007). *Examining Writing: Research and Practice in Assessing Second Language Writing*. Cambridge: Cambridge University Press.

Shepard, L.A. (1993). Evaluating test validity. *Review of Research in Education*, *19*, 405–450.

Sijtsma, K. (2010). Book review. In R.W. Lissitz (Ed.). (2009) The concept of validity: Revisions, new directions, and applications. Charlotte, NC: Information Age Publishing, Inc. *Psychometrika*, *75*(4), 780–782.

Sireci, S.G. (2009). Packing and unpacking sources of validity evidence: history repeats itself again. In R.W. Lissitz (Ed.). *The concept of validity: revisions, new directions, and applications*, (pp.19–37). Charlotte, NC: Information Age Publishing, Inc.

Sireci, S.G. (2012). De-"Constructing" Test Validation. *Center for Educational Assessment Research Report No. 814*. Amherst, MA: Center for Educational Assessment. Paper presented at the annual conference of the National Council on Measurement in Education, Vancouver, April 2012.

Sireci, S.G., Baldwin, P., Martone, A., Zenisky, A.L., Hambleton, R.K., & Han, K. (2006). *Massachusetts Adult Proficiency Tests Technical Manual*. Amherst, MA: Center for Educational Assessment. Retrieved from: http://www.umass.edu/remp/CEA_TechMan.html.

Toulmin, S. (1958/2003). *The uses of argument*. Cambridge: Cambridge University Press.

Text Mining: An introduction to theory and some applications

Nadir Zanini and **Vikas Dhawan** Research Division

Introduction

Recent technological advances have led to the availability of new types of observations and measurements that were previously not available and that have fuelled the 'Big Data' trend (Dhawan & Zanini, 2014). Along with standard *structured* forms of data (containing mainly numbers), modern databases include new forms of *unstructured* data comprising words, images, sounds and videos which require new techniques to be exploited and interpreted.

This article focusses on Text Mining (TM), that is a set of statistical and computer science techniques specifically developed to analyse text data. It aims to give a theoretical introduction to TM and to provide some examples of its applications. Text has always been an informative source of insight into a specific field or individuals. However, with the advent of new technologies, text data are also being predominantly used in new forms of communication. New sources of text data are now available, such as text messaging, social media activity, blogs and web searches. The increasing availability of published text, sophisticated technologies and growing interest in organisations in extracting information from text has led to replacing (or at least supplementing) the human effort with automatic systems.

TM can be used for a variety of scopes, ranging from basic descriptions of text content through word counts to more sophisticated uses such as finding links between authors and evaluating the content of scripts (e.g., automated marking of essays).

TM refers to the process of extracting meaningful numeric indices from text. It owes its origin to a combination of various related fields – Data Mining (DM), Artificial Intelligence, Statistics, Database Management, Library Science and Linguistics (Anawis, 2014). Its basic purpose is to process the unstructured information contained in text data in order to make text accessible to various DM statistical algorithms. This could help make text data as informative as standard structured data and allow us to investigate relationships and patterns which would otherwise be extremely difficult, if not impossible, to discover. With TM, information contained in the text can be categorised and clustered with the aim of producing results such as word frequency distribution, pattern recognition and predictive analytics which might not be easily available using standard data (JISC, 2008).

The possibility of analysing text data is recognised as one of the main elements of the Big Data trend (Lohr, 2012) and a leading source of information for data journalism (Rogers, 2011). In recent years, greater understanding of the potential of TM has led government/public authorities and private organisations to play an active role in developing this technology. The National Centre for Text Mining (NaCTeM) was possibly the first publicly-funded TM centre in the world[1], established by the UK's JISC[2] and operated by the University of Manchester (for an introduction to NaCTeM see Ananiadou, 2005). NaCTeM was established in 2004 to provide TM services in response to the requirements of the UK academic community and to provide leadership in its use in learning, teaching, research and administration. The potential of TM has also been recognised elsewhere in the world. For example, in Italy, Cineca (a consortium made up of 54 Italian universities and the Ministry of Education, University and Research) has been using one of the most powerful computers in the world to design and develop information systems and TM solutions for public administration, health care and business.

TM can be a strategic source of evidence-based information that can support the decision-making process in different fields, from policy-making to business. For this reason, researchers and practitioners from various fields are using TM.

The logic (and technology) behind Text Mining

Broadly speaking, the overarching goal of TM is to turn text into data so that it is suitable for analysis. To achieve this there is a need for applying computationally-intensive artificial intelligence algorithms and statistical techniques to text documents. As stated in a JISC briefing paper (JISC, 2008), TM employs a wide range of tasks that can be combined together into a single workflow, in which it is possible to distinguish four different stages:

1. Information Retrieval
2. Natural Language Processing (NLP)
3. Information Extraction and
4. Data Mining.

Information retrieval

The first stage of TM is to identify the relevant documents from a large collection of digital text documents. Information Retrieval systems used are aimed at identifying the subset of documents which match a user's query. Two examples of Information Retrieval systems are the tools used in libraries to search for books on a specific topic and web search engines (e.g., Google, Bing) designed to search for information in the World Wide Web.

Natural Language Processing

Once a subset of text documents has been retrieved the character strings have to be processed in order to be analysed by computers. The computers need to be fed input in a specific format so that they can understand natural languages as humans do (Manning & Schütze, 1999).

1. See NaCTeM web page at http://www.nactem.ac.uk/
2. JISC (formerly known as the Joint Information Systems Committee) is a UK non-departmental public body whose role is to support post-compulsory education and research, providing leadership in the use of ICT in learning, teaching, research and administration.

The main difficulty is that, often, the hidden structure of natural language is highly ambiguous. Although this might jeopardise the outcome, developments in NLP have led to a high degree of success in certain tasks. NLP enables us to (JISC, 2008):

- classify words into grammatical categories (e.g., nouns, verbs);
- disambiguate the meaning of a word, among the multiple meanings that it could have, on the grounds of the content of the document;
- parse a sentence, that is, perform a grammatical analysis that enables us to generate a complete representation of the grammatical structure of a sentence, not just identify the main grammatical elements in a sentence.

During this stage of TM, the linguistic data about text are extracted from, and marked-up to, the documents which still hold an unstructured form of data.

Information Extraction

In order to be mined as any other kind of data, the unstructured natural language document must be turned into data in a structured form. This stage is called Information Extraction and it is the data generated by NLP systems. The most common task performed during this stage is the identification of specific terms, which may consist of one or more words, as in the case of scientific research documents containing many complex multi-word terms.

Information extraction also allows us to link names and entities (e.g., people and the organisation to which they are affiliated) and more complex facts such as relationships between events or names.

Data Mining

When the structured database is filled with the information extracted from the annotated documents provided by NLP algorithms, data are finally ready to be mined. In this context 'mining' is a synonym of 'analysing', as the aim is to draw useful information from the text data in order to build up new knowledge. To do this, given that data are now in a structured form, it is possible to make use of standard statistical procedures and techniques applied to text data that are now in structured form.[3]

Applications of Text Mining

The first applications of TM surfaced in the mid-1980s.[4] However its growth has been led by technological advances in the last ten years. TM has being increasingly employed in applied research in different areas (such as epidemiology, economics and education) as well as for business-related purposes, especially for gaining market and consumer insights and to develop new products. The techniques of TM are common to both academic research and business-oriented analytics.

3. Among the most common statistical packages used by researchers, the text analytics tools are 'Text Miner' and 'Enterprise Miner' (SAS), 'TM – Text Mining Infrastructure' (R) and 'Modeler' (SPSS).

4. See, for example, the Content Analysis of Verbatim Explanations Research project. http://www.ppc.sas. upenn.edu/cave.htm

From basic word counts to sentiment analyses

Some of the applications of TM require very basic statistics, frequencies for instance. Counting the occurrence of one or more words from a document is the most common TM application, but it does require new ways to visualise this kind of data. For example, Wordle, a free tool available online (http://www.wordle.net/) generates tag clouds of the words contained in a document (Feinberg, 2010). The size of each word is proportional to its relative frequency in the document (similar to a bubble plot).

The technological advances that have fuelled TM development have not just inspired new data visualisations, but also stimulated the collection of new 'textbases', such as *Project Gutenberg* and *Google Books*. For instance, digitising and archiving books allows us to calculate the frequency of a word in a book, or in all the books published in a specific year or to visualise the occurrence of certain words over time. For books available in *Google Books*, Figure 1 gives an example of the occurrence of the words 'information' and 'news' in books published during the last century. Whilst the word 'news' appears to have been steadily used by authors over the last century, the word 'information' experienced a notable increase: from about the same level as 'news' in the early 1900s, to six times more than 'news' in the year 2000.

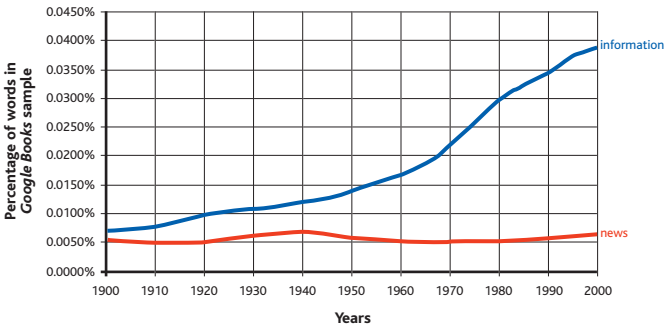


Figure 1: Searches for the words 'information' and 'news' in *Google Books* (digitalised books originally published between 1900 and 2000)

Image sourced from *Google Books Ngram Viewer*.
Retrieved from https://books.google.com/ngrams

Word counts and the availability of large-scale 'textbases' give the opportunity to analyse the evolution of literary styles and trends over time and across countries. This kind of analysis belongs to a new field of study known as 'culturomics' (Ball, 2013). For example, in a recent study, a group of researchers mined a sample of 7,733 works obtained from the *Project Gutenberg* Digital Library written by 537 authors after the year 1550 (Hughes, Foti, Krakauer,& Rockmore , 2012). They focused on the use of 307 content-free words (e.g., prepositions, articles, conjunctions and common nouns) claiming that these words provide a useful stylistic fingerprint for authorship and can be used as a method of comparing author styles. For each author a similarity index with every other author was computed. This index, based on the occurrences of each content-free word considered in the study, was used to exploit temporal trends in the usage of content-free words. Their primary finding was that authors tend to have important stylistic connections to other authors closer to them in time, but not necessarily to immediate contemporaries. They noticed that, for books published within three years of each other, the similarity index is very high, but slightly smaller than the one shown for books published within ten years of each other. For books published with a temporal distance of more than ten years, the similarity index decreased

until reaching a stable value for books published with a temporal distance of 350 years.

Another innovative piece of research, carried out by Matthew Jockers of the University of Nebraska-Lincoln, focused on comparing the stylistic and thematic connections amongst eighteenth and nineteenth century authors. A massive amount of text data using digital versions of nearly 3,500 books was processed to investigate how books were connected to one another on criteria such as frequency of words, choice of words and overarching subject matter (Jockers, 2013). Each book was then affixed with unique attributes and plotted graphically. Figure 2 shows the books analysed from the late 1700s to the early 1900s. The books plotted closer to each other represent a close relationship in terms of styles and themes. Figure 2 highlights the example of Herman Melville's *Moby Dick* published in 1851 which appears here as an outlier from much of the literary work of the period while still being related to several works by James Fenimore Cooper (*Sea Lions* published in 1849 and *The Crater* published in 1847).

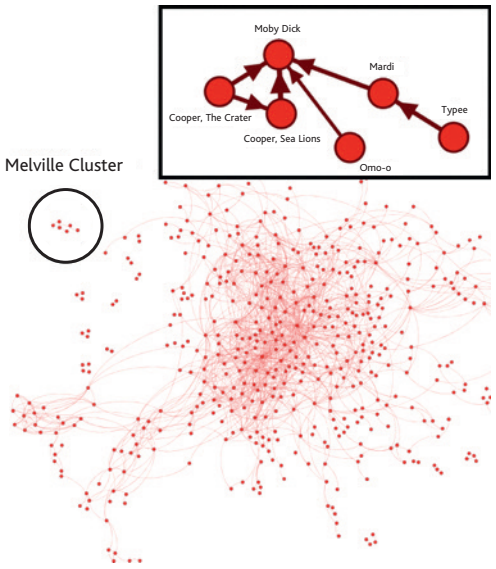


Figure 2: Graphical distribution that displays connections, insights and trends about the literary world from the late 1700s to the late 1900s

Image courtesy of Matthew Jockers (University of Nebraska-Lincoln).

Recent research led by Durham University studied the use of emotion-related words in recent history (Acerbi, Lampos, Garnett & Bentley, 2013). Based on these words this research found that there was a 'sad' peak during the Second World War and two 'happy' peaks – one in the 1920s and another in the 1960s (see Figure 3). A 'sad' period was also noticed during the 1970s and the 1980s followed by an increase in happiness-related words around 1990–2000. The study pointed out that in general, the use of emotion-related words has reduced in the past century. The study also compared historical trends in the use of emotion-related words between British and American authors. Prior to 1980, the difference between them was barely significant, but since then emotion-related words have been used more frequently by US authors than UK ones.

Mining of social opinions is becoming a common marketing and brand management strategy used by organisations. This kind of analysis includes understanding what people say or share in their everyday life, particularly online. This area of research is known as 'opinion mining' or 'sentiment analysis'. Its aim is to identify and extract subjective

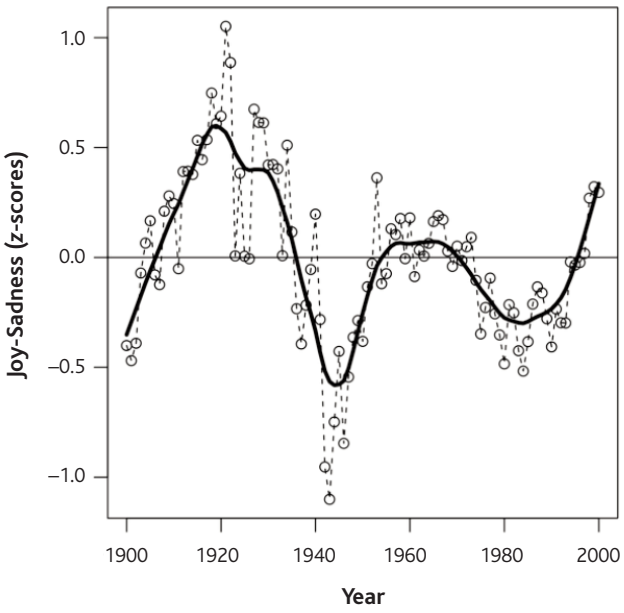


Figure 3: Historical periods of positive and negative moods

Note: Difference between z-scores of Joy and Sadness for years from 1900 to 2000 (raw data and smoothed trend). Values above zero indicate generally 'happy' periods, and values below the zero indicate generally 'sad' periods.

information from text documents such as social media posts. Sentiment analysis is one of the main research strands of Global Pulse, a new initiative by the United Nations (UN) aimed at leveraging the use of Big Data for global development. In a recent work, Twitter conversations related to food price inflation amongst Indonesians were investigated. The research found a significant correlation between official food inflation rates and the number of tweets about this topic (UN, 2014). The study concluded that automated monitoring of public sentiment on social media, combined with contextual knowledge, has the potential to be a valuable real-time alternative to official statistics (usually released after a certain time lag) and to uncover people's reactions in contexts where the use of social media is widespread.

Sentiment mining has also been exploited in other research contexts, such as the understanding of political and historical trends (Ceron, Curini, Iacus & Porro, 2014; Huijnen, Iaan, de Rijke & Pieters, 2014). Social media websites and other computational tools (e.g., *Google Books Ngram Viewer*) are being used for research in this area. This approach could help retrieve hidden information in a large corpus of text documents including speech transcripts by writers and speakers.

Links amongst words and text pattern recognition

Basic statistics are sufficient to summarise, categorise and cluster information from text documents. TM, in addition, may be helpful to generate meaningful links across different documents when decision-makers are overloaded with unstructured information, such as news articles in the case of financial market agents. At times, TM could help reveal unexpected connections between documents. The relationship between the use of certain words in real estate advertisements and the price of the house advertised make an interesting example. In their best-selling book *Freakonomics*, Dubner and Levitt (2005) listed five terms commonly used in real-estate advertisements in the USA associated with

a) higher sale price and b) lower sale price. Table 1 gives the five terms for both (in order of their association with price). The more expensive houses were described using words which were all related to the physical description of the house such as 'granite' and 'maple '. Unexpectedly, words such as 'fantastic' and 'charming' were used more often for cheaper houses. The authors suggest that these words are used as a sort of real-estate agent code to attract potential customers for a house which doesn't have many saleable attributes.

Table 1: Terms used in USA real-estate adverts and their association with house price (Dubner & Levitt, 2005).

Five terms associated with higher price	*Five terms associated with lower price*
Granite	Fantastic
State-of-the art	Spacious
Corian®	!
Maple	Charming
Gourmet	Great neighbourhood

However, we need to be careful in drawing interpretations from text data. For instance, it has been reported in a post written in a language blog by the computational linguistic Mark Liberman[5] that statistically significant correlations were unexpectedly found between words apparently not linked, such as 'some' and 'all', 'the' and 'you'. This suggests that, although it is not hard to find patterns in large datasets, the results may not be meaningful or not always straightforward to interpret and the patterns could also be attributed purely to sampling error.

Word pattern recognition has also been applied to everyday working life. Automated systems (known as eCRM – Customer Relationship Management) have been developed as an attempt to categorise incoming email, and to automatically respond to users with standard answers to frequently asked questions.

One of the most familiar applications of TM technology and machine learning techniques is *Google Translate*, a free, multilingual translation service provided by *Google Inc*. to translate written text from/into 63 languages. *Google Translate* is based on a large scale statistical analysis, rather than traditional grammatical rule-based analysis. To generate a translation, *Google Translate* looks for patterns in hundreds of millions of documents that have already been translated by human translators and are available on the web. This process of seeking patterns in large amounts of text is called 'statistical machine translation' (Och, 2005).[6] Clearly, the more human-translated documents that *Google Translate* can analyse in a specific language, the better the translation quality will be.

Publicly available data and predictive modelling

With the advent of new technologies, a source of data is not just a document for TM, the search for that document itself can provide useful insights. In the case of documents available online, web searches through search engines can be informative. *Google*, for example, set up *Google Trends*, which allows internet users to easily access metrics on *Google* searches.

An example of such trends is given in Figure 4. It shows the comparison of text searches in *Google* for the terms 'OCR', 'Edexcel' and 'AQA' (the names of three awarding bodies based in England, Wales and Northern Ireland) from January 2011 to September 2014.[7] The searches for the three awarding bodies follow a similar pattern to each other which, not unexpectedly, depict a seasonal component: the two peaks are in June and January of year each (except for January 2014[8]), when the majority of students sit the exams, whilst August has fewer searches, when schools are closed. During examination sessions AQA was the most searched, while OCR had the highest number of searches from September to December.[9] *Google Trends* also provides a list of related searches, that is, popular search terms that are associated with the term searched. In the example given here, for all three awarding bodies, the most related search was their name followed by the term 'past papers' (e.g., 'OCR past papers'). The second most frequent related search was the name of the awarding body followed by 'GCSE' (e.g., OCR GCSE). We also observed that while the most searched subject for OCR and AQA was Biology, for Edexcel it was Mathematics.

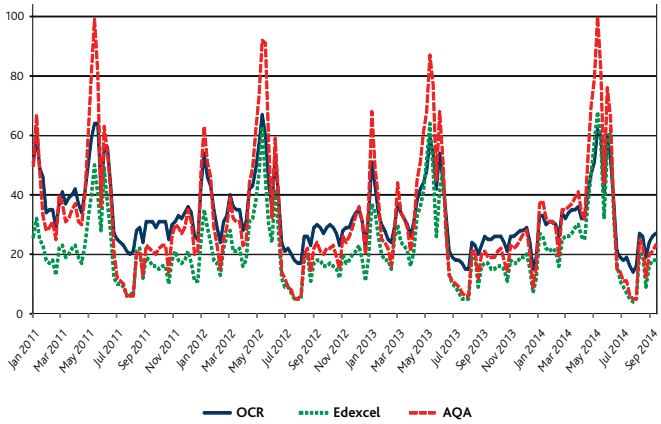


Figure 4: Text searches for 'OCR', 'Edexcel', and 'AQA' from January 2011 to September 2014

Image sourced from *Google Trends*. Retrieved from http://www.google.com/trends

It has been shown that the number of text queries that users enter into web search engines such as *Google* and *Yahoo* can be used for predictive modelling for forecasting values of a number of measures of interest. Researchers in epidemiology discovered that search requests for terms like 'flu symptoms' and 'flu treatments' were a good predictor of the number of patients who, in the period 2004–2008, required access to USA hospital emergency rooms in the next two weeks (Polgreen, Chen, Pennock, Nelson & Weinstein, 2008; Ginsberg et al., 2009). With reference to 2013, it was reported that these web searches were predicting more than double the proportion of doctor visits for influenza-like illness that were actually recorded. This was probably caused by a change in the *Google* search algorithm (Lazer, Kennedy, King, & Vespignani, 2014). Although this discovery can undermine the suitability of web searches as a predictive method, it has been proven to be a good source of

5. *Significant (?) relationships everywhere*. Language Log. Retrieved from: http://languagelog.ldc.upenn.edu/nll/ ?p=4686#more-4686
6. See also the webpage of the Google Research team at http://research.google.com/pubs/MachineTranslation.html
7. *Google Trends* does not provide data on the access to the website (which is something that *Google Analytics* does, though this is not publicly accessible). So the data plotted in Figure 4 are not 'visits' to the three awarding bodies' websites, but only 'searches'. Moreover, data provided does not show the actual volume of searches, but only an indicator estimated in relation to the maximum value of searches across the comparison which is set to 100.
8. It should be noted that in 2014, there was no January exam sitting.
9. Note that the results might have been different if, for instance, 'Pearson' or 'Pearson Edexcel' had been used instead of 'Edexcel'. Pearson has been the parent company of Edexcel since 2003. In 2010, the legal name of the Edexcel awarding body became Pearson Education Limited (Pearson).

information when combined with traditional sources of data. Web search data combined with official statistics have been extensively used to predict the unemployment rate in different countries such as the US (Ettredge, Gerdes, & Karuga, 2005; D'Amuri & Marcucci, 2010), Germany (Askitas & Zimmermann, 2009) and Israel (Suhoy, 2009). It has also been shown that web search data employed as an explanatory variable, along with the previous historical trends of the dimension of interest, can sensibly improve short-term predictions of other social and economic indicators such as inflation (Guzman, 2011). Therefore, predictive modelling could also enable central banks and other national and international agencies to improve the timing and the accuracy of the policy measures they publish to inform policy makers. It can also be applied to economic metrics for business-related purposes and analysing customer insights.

Evidence has shown that web search queries "…can be useful leading indicators for subsequent consumer purchases in situations where consumers start planning purchases significantly in advance of their actual purchase decision" (Choi & Varian, 2012). For instance, search engine data related to housing search enquiries has been shown to be a more accurate predictor of house sales in the next quarter than the forecasts provided by real estate economists (Wu & Brynjolfsson, 2013). Web search queries have also been successfully employed to improve the predictability of motor vehicle demand and holiday destinations (Choi & Varian, 2012). These are applications of the terms attributed to Choi and Varian's – 'contemporaneous forecasting' or 'nowcasting', because they can help in 'predicting the present', rather than the future (Choi & Varian, 2012).

The use of predictive modelling has also been adapted by online retailers to gain customer insights. Amazon and Netflix recommendations, for example, rely on predictive models of what book or film a customer might want to purchase on the basis of their history of enquiries to the website or similar purchases made by other customers (Einav & Levin, 2014). In general, online advertising and marketing tends to rely on automated predictive algorithms that target customers who might be interested in responding to offers.

Predictive modelling based on text data extends well beyond the online world. One of the most famous applications is the development of algorithms that make use of text data contained in different forms of communication (e.g., mobile texts and emails) to detect terrorist threats and to identify fraudulent behaviour in healthcare and financial services (Einav & Levin, 2014).

Applications of Text Mining in education

The benefits offered by the interaction of text and other data analytics in improving learning processes are already being valued by education practitioners as well as by learners themselves.

The first example is the implementation of an experimental real-time case study in a business course. Lecturers made use of internet-based software to facilitate written communication among students, teachers and the case organisation. In this way, it was possible to gather a large quantity of text data containing all the email communication among students and the organisation involved in the case study. Applying simple text analytics on real-time written communication, such as counting of specific words, researchers found that, by the end of this experimental teaching approach, students had increased their understanding of a live business problem. Furthermore, from the analysis of text data, it was possible to discover that, during the case study, students learnt how to use a language more similar to the one used in the real business world. In an evaluation of this experiment, students affirmed that they liked this new teaching approach and would like to see more of it at their schools as they found it very applicable to real life (Theroux, 2009).

A second example of the use of TM to gather insights on learners' cognition is a study aimed at analysing students' progression in a computer programming class. In this study, a software package was used to gather data during a programming assignment from nine learners (Blikstein, 2011). The software allowed researchers to build a 1.5 GB dataset of 18 million lines of events (such as keystrokes, code changes, error messages and actual coding snapshots). An in-depth automated exploration of each student's coding strategies summarised by this mixture of structured and text data was compared with those of other students. The author discovered that error rates progressed in an 'inverse parabolic shape'. This means that, initially, students made a lot of mistakes, but they demonstrated that they were able to learn from them through problem-solving and progressed until they had completed their assignment. Although this is a small-scale study and it is not possible to make any claims about statistical significance, it suggests that using a sophisticated TM application might lead to a better understanding of students' coding styles and sophisticated skills such as problem-solving.

An extensive use of the recent developments in NLP has also been employed to automatically detect secondary students' mental models in order to gain a better understanding of their learning processes. In an experiment students were asked to write short paragraphs about the human circulatory system in order to recall knowledge about the topic. Using an intelligent tutoring system (*MetaTutor*) that teaches students self-regulatory processes during learning of complex Science topics and applying TM techniques, researchers explored which particular machine learning algorithm would enable them to accurately classify each student in terms of their content knowledge (Rus & Azevedo, 2009). Mental models represent an expanding field of research among cognitive psychologists and are aimed at better understanding how well an individual organises content in meaningful ways. TM allows researchers to undertake analysis that can reveal inaccuracies and omissions that are crucial for deep understanding and application of course material, thus informing improvements in course design.[10]

A number of systems using TM have been developed for automated marking of essays and short, free text responses (for an example of the latter see Sukkarieh et al., 2003). Some of the most widely used automated essay marking systems available in the market include: Project Essay Grader, Intelligent Essay Assessor, E-rater, Criterion, IntelliMetric, MY Access and Bayesian Essay Test Scoring System. They have been developed to reduce time and cost and improve reliability and generalisability of the process of assessment in low-stakes classroom tests, as well as for large-scale assessment such as national standardised examinations. The accuracy and reliability of these automated systems have been investigated by educational researchers in the last fifteen years. Along with the benefits of using TM, some of its disadvantages such as the lack of human interaction and the need for a large corpus of sample texts to train the system, have also been reported (Dikli, 2006). Automated essay marking systems do not really understand the texts as

10. For more details on mental model assessment in education see http://mentalmodelassessment.org/

humans do, so it is not possible to affirm that they emulate the human marking process. Notwithstanding, automated essay marking systems show high agreement rates with human markers; and their supporters advocate that the main role of these systems today is not to replace teachers and assessors, but to assist them, incorporating these systems as a supplementary marker, especially in large-scale writing assessments (Monaghan & Bridgeman, 2005; Kersting, Sherin & Stigler, 2014).

A particular example of automated essay marking is the tool developed by a team of researchers at Maastricht University to stimulate students to become active and collaborative learners. It has been used in Statistics courses to assess students on their understanding of course content. It makes use of advanced NLP and Latent Semantic Analysis algorithms that can be used in automatic marking of the texts. Mining students' essays, researchers were easily able to automatically discriminate between the reference book chapter text and the documents of the students. However, it is less clear whether this tool is able to discriminate students from one another (Imbos & Ambergen, 2010).

Despite its weaknesses, marking essays automatically continues to attract the attention of schools, universities, assessment organisations, researchers and educators. Although it might be difficult for these systems to supersede human markers, TM can be employed to support human markers as a second or third marker (see, for instance, Landauer, 2003 and Attali & Burstein, 2006). The Centre for Digital Education (CDE) reported that, in the USA, around $20 billion was spent on public education in Information Technology in 2012, with an increase of 2 per cent from the previous year[11]. The awareness of the potential of TM and DM in, for instance, formative assessment, has led McGraw-Hill to develop two different tools, *Acuity Predictive Assessment* and *Acuity Diagnostic Assessment*, aimed at informing teachers and learners about their performance and how to improve it (CDE, 2014).

These tools can be employed for formative assessment. Predictive modelling of text data can provide an early indication of how students will perform on a standardised test. It allows assessment of the gap between what students are *expected* to know and what they *actually* know. It can also provide evidence regarding which area of the syllabus they have to focus on to improve their performance (West, 2012). Also, more advanced analysis could be informative to teachers about which particular teaching techniques are more efficient for specific students and the best ways to tailor the learning approach to them (Bienkowski, Feng & Means, 2012).

Students' reading comprehension, for example, has been the object of a study based on the use of intelligent tutoring software. The analysis of data such as students' reading mistakes and word knowledge gathered through a speech recognition tool showed that re-reading an old story helped pupils learn half as many words as reading a new story (Beck & Mostow, 2008). An online tool called *WebQuest* provides activities designed for teachers to train pupils in skills such as information acquisition and evaluation of online materials. Students who have experienced these kinds of activities have reportedly enjoyed the collaborative and interactive nature of the activities (Perkins & McKnight, 2005).

Predictive modelling in educational assessment has been mainly based on numeric data (e.g., days of truancy, overall grades and disciplinary problems). However, text data could be used to enable more in-depth analyses in order to get better insights on assessment. For example, Worsley & Blikstein (2011) examined students' dialogues along with other qualitative and quantitative data to develop predictors for student expertise in the area of Engineering design. By leveraging the tools of machine learning, NLP, speech analysis and sentiment extraction, the authors identified a number of distinguishing factors of learners at different levels of expertise. According to the study, these kinds of findings motivate further research in this field and the development of a new paradigm for the evaluation of learner knowledge construction.

11. Centre for Digital Education: http://www.centerdigitaled.com/research/

Discussion

The key advantage provided by TM is the opportunity to exploit text records, on a very large scale. In this article we have briefly described the techniques of TM and some of its applications.

TM has a variety of potential applications in the field of education. In formative and summative assessment, for instance, it could be used to understand trends in vocabulary usage over time and the use of spelling and punctuation. To date, these applications have been carried out by teachers and assessment experts without using advanced techniques such as TM, but TM allows the possibility of implementing these applications on a more comprehensive scale. The developments in NLP allow educational professionals to analyse the language structure of a vast amount of text documents in just a few minutes, plus the ongoing developments in this field could result in an increase in the accuracy of the findings.

The availability of novel data could lead, at least in principle, to novel measurement and research designs to address old and new research questions. However, working with very large, rich and new kind of datasets, it might not be straightforward to figure out what questions the data could answer accurately. Asking the right question might be more important now than ever (Einav & Levin, 2014). Exploiting large text datasets without a proper research question might lead to a significant waste of resources.

More heterogeneous and in-depth data could allow researchers to move from methods that allow the estimation of average relationships in the population towards differential effects for specific subpopulations of interest. This could mean looking at particular categories of students, defined by their specific background, level of achievement and other characteristics of interest. TM is an expanding field with the potential to support innovative areas of research. With careful research designs and proper methods, TM could make a salient contribution to educational research.

References

Acerbi, A., Lampos, V., Garnett, P., & Bentley, R. A. (2013). The expression of emotions in 20th century books. *PloS one*, *8*(3), e59030.

Ananiadou, S., Chruszcz, J., Keane, J., McNaught, J., & Watry, P. (2005). The National Centre for Text Mining: Aims and Objectives. *Ariadne*, *42*. Retrieved from: http://www.ariadne.ac.uk/issue42/ananiadou.

Anawis, M. (2014). Text Mining: The Next Data Frontier. *Scientific Computing*. Retrieved from: http://www.scientificcomputing.com/blogs/2014/01/text-mining-next-data-frontier.

Askitas, N., & Zimmerman, K. F. (2009). Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly* (formerly: *Konjunkturpolitik*), Duncker & Humblot, Berlin, *55*(2), 107–120.

Attali, Y. & Burstein, J. (2006). Automated Essay Scoring With e-rater® V.2. *The Journal of Technology, Learning, and Assessment*, *4*(3).

Ettredge, M., Gerdes, J., & Karuga. G. (2005). Using web-based search data to predict macroeconomic statistics. *Communications of the ACM, 48*(11), 87–92.

Ball, P. (2013, 21 March). Text mining uncovers British reserve and US emotion. *Nature*. Retrieved from: http://www.nature.com/news/text-mining-uncovers-british-reserve-and-us-emotion-1.12642.

Beck, J., & Mostow, J. (2008). How Who Should Practice: Using Learning Decomposition to Evaluate the Efficacy of Different Types of Practice for Different Types of Students. In B. Woolf, E. Aïmeur, R. Nkambou & S. Lajoie (Eds.), *Intelligent Tutoring Systems,* (5091), 353–362. Springer Berlin Heidelberg.

Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief*. U.S. Department of Educational, Office of Educational Technology. Retrieved from: http://tech.ed.gov/wp-content/uploads/2014/03/edm-la-brief.pdf

Blikstein, P. (2011). *Using learning analytics to assess students' behavior in open-ended programming tasks*. Paper presented at the Proceedings of the 1st international conference on learning analytics and knowledge.

Centre for Digital Education (CDE) (2013). Big Data, Big Expectations. *The Promise and Practicability of Big Data for Education*. The Centre for Digital Education. Retrieved from: http://www.centerdigitaled.com/paper/259374351.html

Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New media and society*, *16*(2), 340–358.

Choi, H., & Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record*, *88*(1), 2–9.

D'Amuri, F., & Marcucci, J. (2010). *"Google it!" Forecasting the US unemployment rate with a Google job search index*. ISER Working Paper Series 2009–32. Institute for Social & Economic Research (ISER).

Dhawan, V., & Zanini, N. (2014). Big data and social media analytics. *Research Matters: A Cambridge Assessment Publication*, *18*, 36–41.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, *5*(1).

Dubner, S. J., & Levitt, S. D. (2005). *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*. New York City: William Morrow.

Einav, L., & Levin, J. D. (2014). The Data Revolution and Economic Analysis. *Innovation Policy and the Economy*, *14*(1), 1–24.

Feinberg, J. (2010). Wordle, in J. Steele & N. Iliinsky (Eds.) *Beautiful visualization*, Sebastopol: O'Reilly Media, Inc.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014.

Guzman, G. (2011). Internet search behavior as an economic forecasting tool: The case of inflation expectations. *Journal of Economic and Social Measurement*, *36*(3), 119–167.

Hughes, J.M., Foti, N. J., Krakauer, D. C., & Rockmore D. N. (2012). Quantitative patterns of stylistic influence in the evolution of literature. *Proceedings of the National Academy of Sciences*, *109*(20), 7682–7686.

Huijnen, P., Laan, F., de Rijke, M., & Pieters, T. (2014). A Digital Humanities Approach to the History of Science. In A. Nadamoto, A. Jatowt, A. Wierzbicki & J. Leidner (Eds.), *Social Informatics,* (8359), 71–85. Springer Berlin Heidelberg.

Imbos, T., & Ambergen, T. (2010). *Text analytic tools for the cognitive diagnosis of student writings*. Paper presented at the Proceedings of the ICOTS8, International Conference on Teaching Statistics.

JISC (2008). *Text Mining Briefing Paper*. Joint Information Systems Committee. Retrieved from: http://jisc.ac.uk/media/documents/publications/bptextminingv2.pdf.

Jockers, M. (2013). *Macroanalysis: Digital Methods and Literary History*. Champaign: University of Illinois Press.

Kersting, N. B., Sherin, B. L. & Stigler, J. W. (2014). Automated Scoring of Teachers' Open-Ended Responses to Video Prompts: Bringing the Classroom-Video-Analysis Assessment to Scale. *Educational and Psychological Measurement*, *74*(6), 950–974.

Landauer, T. K. (2003). Automatic Essay Assessment, Assessment. *Education: Principles, Policy & Practice*, *10*(3), 295–308.

Lazer, D., Kennedy, R., King, G. & Vespignani, A. (2014, 14 March). The Parable of Google Flu: Traps in Big Data Analyisis. *Science*, *343*(6176), 1203–1205. Retrieved from: http://www.sciencemag.org/content/343/6176/1203.

Lohr, S. (2012, 11 February). The Age of Big Data. *The New York Times*. Retrieved from: http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all&_r=0.

Manning, C.D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Boston: MIT press.

Monaghan, W., & Bridgeman, B. (2005). E-rater as a Quality Control on Human Scorer. *ETS RD Connections*. Retrieved from: http://www.ets.org/Media/Research/pdf/RD_Connections2.pdf.

Och, F.J. (2005). *Statistical Machine Translation: Foundations and Recent Advances*. Retrieved from: http://www.mt-archive.info/MTS-2005-Och.pdf.

Perkins, R., & McKnight, M.L. (2005). Teachers' attitudes toward WebQuests as a method of teaching. *Computers in the Schools*, *22*(1–2), 123–133.

Polgreen, P.M., Chen, Y., Pennock, D. M., Nelson, F. D., & Weinstein, R. A. (2008). Using Internet Searches for Influenza Surveillance. *Clinical infectious diseases*, *47*(11), 1443–1448.

Rogers, S. (2011, 28 July). Data journalism at the Guardian: what is it and how do we do it? *The Guardian Datablog*. Retrieved from: http://www.theguardian.com/news/datablog/2011/jul/28/data-journalism

Rus, V., Lintean, M., & Azevedo, R. (2009). *Automatic Detection of Student Mental Models during Prior Knowledge Activation in MetaTutor*. International Working Group on Educational Data Mining. Paper presented at the International Conference on Educational Data Mining (EDM) (2nd, Cordoba, Spain, July 1–3, 2009).

Sukkarieh, J. Z., Pulman, S. G. & Raikes, N. (2003). *Auto-marking: using computational linguistics to score short, free-text responses*. Paper presented at the Proceedings of 29th International Association for Educational Assessment (IAEA) Annual Conference.

Suhoy, T. (2009). *Query indices and a 2008 downturn: Israeli data*. Discussion paper No. 2009.06. Research Department, Bank of Israel.

Theroux, J.M. (2009). Real-time case method: analysis of a second implementation. *Journal of Education for Business*, *84*(6), 367–373.

United Nations (UN) (2014). *Mining Indonesian Tweets to Understand Food Price Crises*. UN Global Pulse Report. Retrieved from: http://www.unglobalpulse.org/sites/default/files/Global-Pulse-Mining-Indonesian-Tweets-Food-Price-Crises%20copy.pdf.

West, D.M. (2012). *Big Data for Education: Data Mining, Data Analytics, and Web Dashboards*. Retrieved from: http://www.brookings.edu/research/papers/2012/09/04-education-technology-west.

Worsley, M., & Blikstein, P. (2011). *Using machine learning to examine learner's engineering expertise using speech, text, and sketch analysis*, in Paper presented at the 41st Annual Meeting of the Jean Piaget Society (JPS). University of California, Berkeley.

Wu, L., & Brynjolfsson, E. (2013). The future of prediction: How Google searches foreshadow housing prices and sales, in S. M. Greenstein, A. Goldfarb and C. Tucker (Eds.) *Economics of Digitization*, Chicago: University of Chicago Press.

Research News

Karen Barden Research Division

Conferences and seminars

Cambridge Assessment's Conference 2014

The seventh Cambridge Assessment Conference took place in October with the theme of International Education: Interpretation, Importance and Impact. We welcomed over a hundred industry leaders from fifteen countries who gathered in Cambridge to discuss the challenges and opportunities that education without borders creates, and to explore its different facets and impact on local governments, training providers and students worldwide.

An understanding of an international education in the context of global issues was presented by keynote speaker David Smith, Economics Editor, The Sunday Times. An interpretation of an international education was discussed by Marc Tucker, President and CEO of the National Center on Education and the Economy, and Dr David Graddol, Director of The English Company (UK) Limited. Isabel Nisbet of the A Level Content Advisory Board, Professor Jeremy Hodgen, Professor of Mathematics Education, King's College London and Dr Karin Zimmer, Researcher, German Institute for International Pedagogical Research, each considered the importance of an international education and the improvement of national systems.

Case studies from around the world were explored by Dr Stephen Burr, Managing Director of Reddam House Europe; David Barrs, Head Teacher at the Anglo European School; Dino Varkey, Group Executive Director and Board Member, GEMS Education and Gisella Langé, Foreign Languages Inspector, Italian Ministry of Education.

The most unique aspect of the day was the contributions via video-links and social media from some of our schools in India, Egypt, South Africa, Argentina and Mexico. Having contributors from so many different countries and time zones made it a challenging programme but the technology all worked on the day and there was much debate about how international education will continue to grow and that it has something positive and distinctive to contribute.

Further details, podcasts and an opportunity to share your story on what international education means to you, can be found on our website: www.cambridgeassessment.org.uk/conference2014/

Kaleidoscope Graduate Student Research Conference

The Kaleidoscope Research Conference was held at the Faculty of Education, University of Cambridge in May. Magda Werno presented a paper on *Entering secondary education in England as a non-native speaker: A case study of transitional experiences and initial support*.

Professional Practice, Education and Learning (ProPEL) Conference

The Second International ProPEL Conference took place in Stirling, Scotland in June. It provided an opportunity to debate leading edge studies in professional and vocational learning, practice and education. Martin Johnson presented a paper entitled *The work of making it happen*.

European Association for Research on Learning and Instruction (EARLI) – SIG 4: Higher Education

This conference was held in Leuven, Belgium in August and explored the theme of Assessing Transitions in Learning. Tom Sutch presented a paper on *The effect of specialism and attainment in secondary school on the choice of HE institution and field of study*.

European Association for Research on Learning and Instruction (EARLI) – SIG 1: Assessment & Evaluation

In August, Frances Wilson and Jackie Greatorex attended the EARLI SIG 1 Conference in Madrid, Spain. The main themes were teachers' assessment literacy, professional learning communities and requirements for professional development. Frances presented a paper entitled *Teachers' use of differentiated assessment: the tiering model*. Jackie presented two papers; one on *Context in Maths exam questions* and the second entitled *Around the world in Cambridge International A Level Mathematics: teachers' views*.

European Conference on Educational Research (ECER)

Held in Porto, Portugal in September, the ECER Conference provided an opportunity to debate The Past, the Present and Future of Educational Research. Carmen Vidal Rodeiro presented a paper on *Academic and vocational pathways to higher education and their impact on the choice of institution and field of study*. Simon Child's paper presentation was entitled *Framing educational change: Teacher and employer voices in the development of new courses in English for 16-year-olds*.

British Educational Research Association (BERA)

BERA celebrated the 40th anniversary of its Annual Conference by exploring the advances made in educational research since 1974. The conference was held from 23–25 September at the Institute of Education, University of London. Colleagues from the Research Division presented the following papers:

Tom Benton: *The relationship between time in education and achievement in PISA in England*.

Ellie Darlington: *Differences between A level and undergraduate Mathematics questions: doing, reproducing or practising Mathematics* and *Shortcomings of the approaches to learning framework in the context of undergraduate Mathematics*.

Tim Gill: *An investigation of the effect of early entry on overall GCSE performance, using a Propensity Score Matching method*.

Jackie Greatorex: *Context in Maths exam questions*.

Nicky Rushton and Frances Wilson: *Teachers' and employers' views on the transition from GCSE Mathematics to A level Mathematics and employment*.

Carmen Vidal Rodeiro: *Progression routes to post-16 Science qualifications*.

Carmen Vidal Rodeiro, Tom Sutch and Nadir Zanini: *Pathways to Higher Education: the effect of different prior qualifications on institution and field of study*.

Frances Wilson and Vikas Dhawan: *Capping of achievement through tiering at GCSE*.

Nadir Zanini: *How do A level subjects and grades determine university choices?*

Tom Benton also presented a poster on *Evaluating the reliability of PISA using simple methods*.

Association for Educational Assessment – Europe (AEA-Europe)

The 15th AEA-Europe Annual Conference took place in Tallinn, Estonia in November with the theme of Assessment of students in a 21st Century world. Several colleagues from Cambridge Assessment attended the conference and the following papers were presented:

Tom Bramley: *Evaluating the 'adjacent levels' model for differentiated assessment*.

Victoria Crisp: *Judgement in the assessment of 'harder to examine' skills: what do assessors pay attention to?*

Victoria Crisp and Stuart Shaw: *Evaluating assessments in the 21st Century: Reflections on a framework for validation – 5 years on*.

Martin Johnson and Beth Black: *How do examiners align their marking judgements remotely? Insights into examiners' feedback interactions for remote standardisation*.

Tim Oates: *Can 21st Century science qualifications exist without assessment of practical science skills?*

Stuart Shaw and Paul Newton (Institute of Education, University of London): *21st Century Evaluation: towards a neo-Messickian framework for the evaluation of testing policy*.

The following posters were also presented:

Sarah Hughes and Stuart Shaw: *To remark or review? Modelling examiner behaviour*.

Stuart Shaw and Paul Newton (Institute of Education, University of London): *Tracing the trajectory of the evolution of validity: key phases in the history of validity theory*.

.

Society for Research into Higher Education (SRHE)

The 2014 SRHE conference was held in Newport, Wales in December. The conference explored the theme of Inspiring future generations; embracing plurality and difference in higher education. Nadir Zanini presented a paper on *The role of the A* grade at A level as a predictor of university performance*.

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

Publications

The following articles have been published since Issue 18 of *Research Matters*:

Benton, T. Examining the impact of entry level qualifications on educational aspirations. *Educational Research*, *56*(3), 259–276.

Benton, T. An Empirical Assessment of Guttman's Lambda 4 Reliability Coefficient in Millsap, R.E., Bolt, D.M., van der Ark, L.A., and Wang, W-C (Eds.) (2015). *Quantitative Psychology Research. The 78th Annual Meeting of the Psychometric Society* (pp.301–310). Springer: New York. doi: 10.1007/978-3-319-07503-7_19

Darlington, E. Contrasts in mathematical challenges in A-level Mathematics and Further Mathematics, and undergraduate Mathematics examinations. *Teaching Mathematics and its Applications*, *33*(4), 219–229. doi:10.1093/teamat/hru021

Johnson, M (2014). Insights into contextualised learning: how do professional examiners construct shared understanding through feedback? *E-Learning and Digital Media*, *11*(4), 363–378. Available online at: http://learning1060.rssing.com/browser.php?indx=4980578&last=1&item=5

Johnson, M (2014). A case study of inter-examiner feedback from a UK context: Mixing research methods to gain insights into situated learning interactions. *Formation et pratiques d'enseignement en questions*, *17*, 67–88. Available online at: http://www.revuedeshep.ch/site-fpeq-n/Site_FPEQ/17_files/05-Johnson.pdf

Shaw, S. D., Warren, J. and Gill. T. (2014). Assessing the Impact of the Cambridge International Acceleration Program on U.S. University Determinants of Success: A Multi-Level Modelling Approach. *College and University Educating the Modern Higher Education Administration Professional*, *89*(4), 38–56. Available online at: https://aacrao-web.s3.amazonaws.com/files/yUcGnGuwS0MFcRDvS4gn_CUJ8904-Web.pdf

Suto, I., Elliott, G., Rushton, N., and Mehta, S. Course struggle, exam stress, or a fear of the unknown? A study of A level students' assessment preferences and the reasons behind them. *Educationalfutures, British Education Studies Association*, *6*(2), 21–43. Available online at: http://educationstudies.org.uk/?post_type=journal&p=2748

Warwick P., Shaw, S. D. and Johnson, M. (2014). Assessment for Learning in International Contexts: exploring shared and divergent dimensions in teacher values and practices. *The Curriculum Journal*, *25*(4), 1–31. doi:10.1080/09585176.2014.975732

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 its Business Streams, 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/

Statistical Reports

Tim Gill Research Division

The on-going '*Statistics Reports Series*' provides statistical summaries of various aspects of the English examination system such as trends in pupil uptake and attainment, qualifications choice, subject combinations and subject provision at school. These 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 this series are:

- *Statistics Report Series No.70: Uptake of GCSE subjects 2013*
- *Statistics Report Series No.71: Provision of GCSE subjects 2013*
- *Statistics Report Series No.72: Uptake of GCE A level subjects 2013*
- *Statistics Report Series No.73: Provision of GCE A level subjects 2013*
- *Statistics Report Series No.74: Uptake of Modern Foreign Languages at GCSE 2013*
- *Statistics Report Series No.75: Uptake of GCE AS level subjects 2007–2013.*

Stop Press : CBE for Tim Oates

As this issue of *Research Matters* goes to press, we are delighted to announce that Tim Oates, Group Director of Assessment Research and Development, has been appointed a Commander of the Order of the British Empire (CBE) in the New Year Honours. Tim, who chaired the 2010 National Curriculum review, has advised the UK Government for many years on both practical matters and assessment policy and has been with Cambridge Assessment since May 2006. The award recognises his services to Education.

Tim says he first heard about the honour a few weeks ago but, following usual protocol, was sworn to secrecy until it was officially announced on 30 December 2014. The news of his CBE was reported on the TES website and listed in both *The Times* and the *Guardian* newspapers.

"I am pleased to receive this honour; I would like to thank all those at Cambridge Assessment who enabled me to make this contribution to improving our education system," Tim said.

Simon Lebus, Group Chief Executive of Cambridge Assessment, said: "We are all delighted that Tim has been so honoured; it is signal recognition of the body of work he continues to produce as a Group Director at Cambridge Assessment".

CONTENTS : Issue 19 Winter 2015

2 **A level History choices: Which factors motivate teachers' unit and topic choices?** : Simon Child, Ellie Darlington and Tim Gill

7 **Context led Science courses: A review** : Frances Wilson, Steve Evans and Sarah Old

14 **Assessing active citizenship: An international perspective** : Prerna Carroll, Simon Child and Ellie Darlington

19 **An investigation into the numbers and characteristics of candidates with incomplete entries at AS/A level** : Carmen Vidal Rodeiro

26 **The moderation of coursework and controlled assessment: A summary** : Tim Gill

31 **Reflections on a framework for validation – Five years on** : Stuart Shaw and Victoria Crisp

38 **Text Mining: An introduction to theory and some applications** : Nadir Zanini and Vikas Dhawan

45 **Research News** : Karen Barden

47 **Statistical Reports** : Tim Gill

47 **Cambridge Mathematics launch**

48 **Stop Press : CBE for Tim Oates**

Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU
UK

Tel: +44(0)1223 552666
Fax: +44(0)1223 552700
Email: researchprogrammes@cambridgeassessment.org.uk
www.cambridgeassessment.org.uk

ISSN: 1755–6031