



**Cambridge
Assessment**

News articles as data: analysing the portrayal of exams in the UK print media

Conference Paper Abstract

Matthew Carroll

Presented at the 44th International Association for Educational
Assessment (IAEA) Annual Conference, Oxford, UK,
September 2018

Author contact details:

Matthew Carroll
Assessment Research and Development,
Research Division
Cambridge Assessment
The Triangle Building
Shaftesbury Road
Cambridge
CB2 8EA
UK

Carroll.M@cambridgeassessment.org.uk

<http://www.cambridgeassessment.org.uk>

As a department of Cambridge University, Cambridge Assessment is respected and trusted worldwide, managing three world-class examination boards, and maintaining the highest standards in educational assessment and learning. We are a not-for-profit organisation.

How to cite this publication:

Carroll, M. (2018, September). *News articles as data: analysing the portrayal of exams in the UK print media*. Paper presented at the 44th International Association for Educational Assessment (IAEA) Annual Conference, Oxford, UK.

Abstract

Content analysis of news provides insights into the discourse around current issues. However, the vast quantity of articles produced can make traditional content analysis difficult. In recent years, text mining techniques have been developed that aid extraction of information from large bodies of text, potentially allowing large numbers of news articles to be analysed. Here, such methods were used to examine the portrayal of GCSEs, academic qualifications taken at age 16 in England, Wales and Northern Ireland that receive extensive media attention.

A news database was searched for articles containing the word “GCSE”, published in national newspapers from 1988 onwards (when GCSEs were introduced). Articles were processed using the R statistical program. Duplications, league tables and live blogs were removed, as were articles that did not contain “GCSE” in the title or ≥ 4 times in the text. Common but uninformative words (structural words, newspaper names, etc.) were removed. Remaining words were lemmatized (removing inflections whilst retaining the grammatical form). The final corpus contained over 2 million words from nearly 7,000 articles.

The most common words were identified: these indicated a focus on pupils, exams, results, and on English and maths specifically. Words were assigned ‘positive’ or ‘negative’ sentiments using a sentiment lexicon: common negative words related to failure, difficulty and declines, whilst common positive words related to good results, skills and improvement. Latent Dirichlet allocation was used to automatically identify topics, and the proportion of articles falling into each was assessed: common topics related to results, reforms and performance comparisons. Further analysis explored how key words, sentiments and topic coverage changed over time.

By applying text mining methods to thousands of articles, this research highlights key themes in media coverage of GCSEs. Further, it shows how changes in coverage over time could reflect, or even influence, changing perceptions of exams.