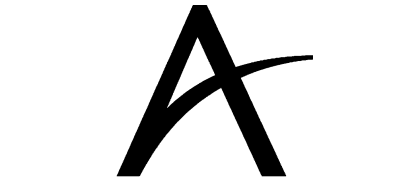

How should grade boundaries be determined in examinations? An exploration of the script features that influence expert judgements

A paper presented at the European Conference for Educational Research, Vienna, Austria, September 2009.

Nadezda Novakovic and Irenka Suto
E-mail: suto.i@cambridgeassessment.org.uk

Abstract

Background: In England and Wales there operates a system of externally designed and assessed qualifications for secondary-school students which affect entry into employment and higher education. Given the high-stakes nature of the examinations involved, there is a strong public expectation that examination standards, expressed through grade boundaries, are rigorously maintained over time. As several European countries are moving from school-based to external assessments, these concerns may well become more prevalent across Europe, increasing the likelihood of international comparisons of examination standards.

The process of determining grade boundaries for many examinations in England relies heavily upon human judgement, in contrast with purely statistical methods. The putative advantage of the judgemental methods is that they allow experts to actually see what students have written in their scripts before passing judgement on the quality of their work. Much research activity is therefore devoted to the scrutiny and development of methods of capturing the judgements of expert examiners. In this paper we focus on three such methods: traditional method, Thurstone pairs and rank-ordering. The traditional method relies on experts making absolute judgements about the quality of students' work. In contrast, the Thurstone pairs and rank-ordering methods require experts to make comparative judgements of students' performances. The recent psychological literature suggests that humans are indeed better at making comparative than absolute judgements (Laming 2004), thus lending theoretical support to these alternative methods.

We investigated the validities of the three methods by exploring the appropriateness of features that judges attend to when making judgements about script quality. The features focussed upon were selected on the basis of references to them in the literature and/or in the grade descriptions for the syllabuses under investigation. Our aim was to identify the features that judges attended to the most in each method.

Methodology: We explored students' scripts from around the lowest and the highest grade boundaries from past examinations for 16-18-year-olds in two contrasting subjects: English and biology.

For each subject a 3x3 'Latin square' design was used, entailing three groups of ten judges and three matched sets of examination scripts. Each judge group made judgements using each of the three methods on a different set of scripts, in a unique order. This design enabled comprehensive comparisons of the three methods, whilst controlling for order effects.

Additionally, every script was rated by two judges independently on nine script features, using rating scales constructed by the researchers. The mean feature ratings were used to determine which features the judges attended to most in each method. For Thurstone pairs and rank-ordering, the mean feature ratings were analysed with multiple regression, while for the traditional method, a simpler quantitative analysis was used.

Conclusions: Few features were identified as being influential in all three methods, and some of the features influenced the judgements in the three methods differently, confirming qualitatively distinct judgemental approaches. The methods also differed in whether judges attended to the same or different features at the highest and lowest grade boundaries. However, the study provided no empirical evidence to support one method over the other based on the features that judges attend to.

We conclude with a summary of key practical and theoretical advantages and disadvantages of each method, to support European decision-makers in choosing methods most suited to their own educational contexts.

Reference:

Laming, D. (2004) Human judgment: the eye of the beholder. London: Thomson.

Keywords: assessment, external examinations, judgemental methods, script features, validity