Multiple marking using the Levels-only method for A level English Literature

Conference Abstract

Jo Ireland

Emily de Groot

AEA-Europe conference, 1-4 November 2023

Author contact details:

Jo Ireland & Emily de Groot
Assessment Research and Development,
Research Division
Shaftesbury Road
Cambridge
CB2 8EA
UK

jo.ireland@cambridge.org
clare.degroot@ocr.org.uk
https://www.cambridge.org/

How to cite this publication:

Ireland, J., & de Groot, E. (2023, November 1-4). *Multiple marking using the Levels-only method for A level English Literature* [Paper presentation]. Annual conference of the Association for Educational Assessment – Europe (AEA-Europe), Malta. https://2023.aea-europe.net/

Essays are important assessment tools allowing us to capture constructs that other assessment items cannot (Holmes et al., 2017). To mark essays, analytical marking methods are commonly used. These can involve examiners allocating marks, nested within levels of performance for different areas of achievement or features of the essay (Meadows and Billington, 2005). However, this process can be time and labour intensive and thus typically only one examiner marks each script. This potentially renders the marking less reliable as it becomes subject to each examiner's individual preference, lenience, or severity, particularly for more subjective subjects such as English. Whilst this can be mitigated with standardisation processes, marker monitoring and statistical scaling of results (e.g., Benton and Gallacher, 2018); more subjective subjects and extended writing tasks still tend to have lower marking reliabilities (Holmes et al., 2017; Wheadon, de Moira et al., 2020).

One approach to improve validity and reliability, which this research explores, is to make essay marking quicker through less detailed marking methods, facilitating multiple marking. In multiple marking, several examiners independently assess the same script, where the final mark is a combination of the individual marks.

In a recent study, Walland and Benton (2021) developed a less time intensive essay marking method called "Levels only (LO) marking" whereby the essay is marked using only the levels for each AO in the existing mark scheme. Marks within the levels are not allocated and no annotations or summative comments are made. Walland and Benton tested it on a single essay from a GCSE English Language component. Their results for both double and triple LO marking were encouraging, yielding high reliability and predictive value. Triple marking used similar examiner time as traditional marking.

This paper looks at how well the LO marking method works with longer A level English Literature essays, and for a whole component. We selected a representative set of 150 scripts from Summer 2019 and recruited a representative set of 10 examiners (principals, senior team leaders, and assistant examiners etc.) who had previously marked A level English. Each examiner marked 60 scripts therefore allowing each script to be marked 4 times.

We found similar encouraging results to Walland and Benton (2021) with high reliability and predictive value.

For reliability, we estimated the variance attributed to the essay only as a proportion of the overall variance in individual LO scores using a mixed-effects linear model. For single marking, we estimate that 68% of the total variance was contributed by the essay, a little lower than the standard benchmark of 70% for "good" reliability. However, the reliability estimates increased to 81% and 86% for double and triple marking respectively. For predictive value, we measured the ability of the two marking methods on the first A level English Literature paper (used in our study) to predict the (original) marks from the second A level English Literature paper (not used in our study). The predictive value (correlation) from the original mark scheme was 0.68. While the predictive value of single marking LO was lower at 0.65, the predictive value increased to 0.71, 0.73, and 0.76 for double marking triple marking, and quadruple respectively.

However, in our study, double marking took a similar amount of examiner time as traditional marking, thus rendering triple marking unfeasible in practice. In particular, the average time to mark an A level English Literature paper in the live exam was 22.9 minutes, whereas the average time of the examiners in this research study was 10.8 minutes, a little under half the traditional time.

Of course, we interpret these results on both reliability and time with caution, recognizing that specific features of the experimental setting and the unfamiliar nature of the new method for examiners widens the confidence bands around estimates.

In addition to our quantitative analysis of the LO method, we also gathered qualitative examiner feedback.

Five respondents mentioned the speed of the method as a positive. The holistic nature of the method was a positive for two markers, while other comments cited the focus on AOs, the ease of use/simplicity, fairness for candidates, and satisfaction with the mark scheme layout as benefits. One examiner said there was nothing they disliked about the method.

The most frequently mentioned dislike was a sense of uncertainty. Reasons for this included lack of familiarity with the new method; lack of standardisation and confidence in decisions; lack of annotations; and concern about fairness of the LO approach. However, there was some recognition that it can take time to adjust to a change in method. Thus, in a real setting, with more training and support, this suggests the results could be even more encouraging.

References

Benton, T., & Gallacher, T. (2018). Is comparative judgement just a quick form of multiple marking. Research Matters: A Cambridge Assessment Publication (26), 22-28.

Holmes, S., Black, B., & Morin, C. (2017). Marking reliability studies 2017: Rank ordering versus marking – which is more reliable? https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/859250/Marking_reliability_-_FINAL64494.pdf

Meadows, M., & Billington, L. (2005). A review of the literature on marking reliability. https://research.aqa.org.uk/sites/default/files/pdf_upload/CERP_RP_MM_01052005.pdf

Walland, E. & Benton, T. (2021). Multiple marking methods as alternatives to analytical essay marking: Comparing pairwise comparative judgement, rank ordering and levels-only. Cambridge Assessment

Wheadon, C., de Moira, A. P., & Christodoulou, D. (2020). *The classification accuracy and consistency of comparative judgement of writing compared to rubric-based teacher assessment*. https://doi.org/10.31235/osf.io/vzus4