Research paper: executive summary

International Baccalaureate
Baccalauréat International
Bachillerato Internacional

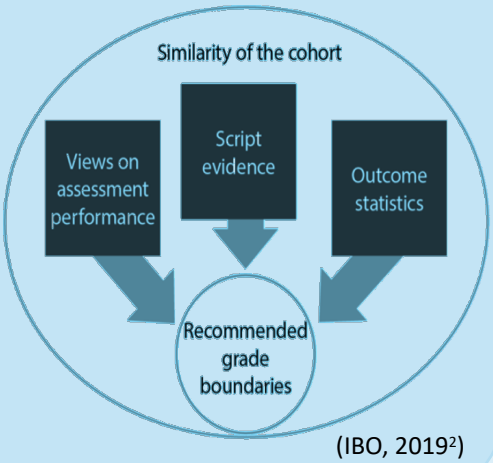
Statistically guided grading judgements: contextualisation or contamination?

Full article: Statistically guided grading judgements: contextualisation or contamination? (tandfonline.com)[1]

INTRODUCTION

Different sources of assessment evidence are reviewed during International Baccalaureate (IB) grade awarding to convert marks into grades and ensure fair results for students. Qualitative and quantitative evidence are analysed to determine grade boundaries, with statistical evidence weighed against examiner judgement and teachers' feedback on examinations. In Spring 2022, the IB conducted a trial to explore how examiners' grading decisions were influenced by having access to statistical evidence and teachers' feedback. The purpose was to compare different approaches to IB grade award, and determine if grading outcomes and experiences varied substantially if examiners were not given access to additional assessment evidence prior to grading scripts.

Evidence that supports the selection of grade boundaries



(IBO, 2019[2])

CONTEXT

In grade award (GA), the IB uses a range of evidence to determine the most suitable grade boundaries for each subject and exam.

Historically, in most subjects, examiners were provided with assessment evidence to review prior to grading scripts (e.g., statistically recommended boundaries (SRBs), mean marks, mark distributions, etc.) in an 'extended model' of GA.

However, in May 2019, a 'limited model' was introduced in some subjects, where examiners graded scripts without access to this additional evidence.

Extended model:
Examiners review statistical data & teacher feedback, then grade scripts.

Limited model:
Examiners grade scripts without access to statistical data and teacher feedback.

AIMS

The overall aim of this study was to compare grading outcomes and examiner experiences in the extended and limited model of GA. This included investigating how different sources of evidence are integrated in IB grade award, and how examiners' grading decisions are impacted by reviewing contextual evidence before making their grading judgements.

The study was guided by the following research questions:

1. How and to what effect are judgemental and statistical evidence combined during IB grade award?
2. To what extent does access to statistical evidence on exams impact examiners' grading decisions?
3. Do grade awards when examiners review statistical evidence lead to similar grading outcomes, compared to when they do not?
4. What are the perceived benefits and drawbacks of examiners reviewing statistical evidence on exams in IB grade award?

[1] Badham, L. (2023). Statistically guided grading judgements: contextualisation or contamination? *Oxford Review of Education,* DOI: 10.1080/03054985.2023.2290640.
[2] IBO, (2019). *Assessment principles and practices – Quality assessments in a digital age.* IBO, Cardiff.

METHODOLOGY

Grade award processes were replicated in nine exams across five subjects. 30 participants took part in the study (25 examiners & 5 subject managers), with all assessment materials taken from May 2019.

Each subject followed the opposite GA model compared to May 2019:

Subject	Extended model	Limited model
English Lit HP1 & HP2	Trial	May 2019
Japanese Lit HP1 & HP2		
Spanish Lit HP1 & HP2		
Business management HP1 & HP2	May 2019	Trial
MYP Maths		

A mixed methods approach was employed, with quantitative comparisons carried out on grading outcomes, and focus groups held to gather feedback on participants' experiences.

FINDINGS

Quantitative

Preliminary findings suggest that both approaches lead to broadly comparable grading outcomes. The proportion of scripts judged to be 'grade-worthy' were similarly aligned to the SRBs in both models. The only exception was Japanese literature, where the divergence is likely explained by the very small cohort size & limited availability of scripts for decision-making.

Qualitative

Five main themes emerged from the focus groups:

1. Script evidence is considered central to examiners' role in grade award.
2. Statistical recommendations can cloud examiner judgement during grading.
3. The most useful statistical evidence for guiding examiners' grading is item-level data.
4. Statistical evidence has many uses for examiners outside of grading.
5. One size does not fit all: needs and requirements vary across subjects.

LIMITATIONS

- Small number of subjects: evidence may be combined and valued differently in other disciplinary contexts.
- Potential risk that some participants may have remembered original grade boundaries.

Due to these limitations, the primary aim of the study was to encourage discussion on how best to combine different forms of evidence in GA, rather than to provide definitive evidence in favour of one particular approach.

RECOMMENDATIONS

1. **Item-level data:** continue providing item-level data to examiners in subjects where it is available, to support examiners & subject managers in identifying 'key discriminator' questions in grade boundary setting. Investigate the possibility of sourcing item-level data in subjects/components where it is not currently available (e.g., Language A Paper 2).
2. **Awarding timeline:** to minimise the risk of clouding examiner judgement in grading as well as reducing the admin burden on subject managers, consider sharing other statistical data with senior examiner after the session, rather than prior to grading.
3. **Flexibility by subject:** aim to balance the need for consistent processes with the requirements & demands of individual subjects (e.g., arising from differences in cohort sizes, mark ranges or language of instruction).
4. **Communication with senior examiners:** to avoid examiners feeling disenfranchised, try to communicate final grade boundary decisions – as well as how their input contributed to these, and how the different forms of evidence were combined to ensure the fairest outcomes for students.
5. **Relevance for other examiner tasks:** as assessment evidence is useful for other examiner activities such as paper setting, subject report writing & maintaining marking standards, it may also be useful to share with more senior examiners outside the awarding team.
6. **Further research:** possibilities for further investigations include replicating the study in different subject areas, or exploring the impact of examiners reviewing statistical recommendations post hoc rather than before grading. Investigations into the impact for very small cohort subjects would also be particularly useful, as statistical evidence is inherently less reliable in these contexts.