Pearson

Pearson School Assessment and Qualifications at AEA-Europe 2025

The Hague, Netherlands

Table of Content

Presenters 3

Posters 4

Test Equating for maintaining standards over time in similar qualifications using Comparative Judgement **4**
Kevin Mason 4

Theoretical Framework for Digital-First Assessment **5**
Dr Liyuan Liu, Irene Custodio 5

Using longitudinal datasets to explore labour market outcomes for learners taking Vocational and Technical Qualifications in England **6**
Blake Ashworth, Hayley Dalton, Ria Bhatta 6

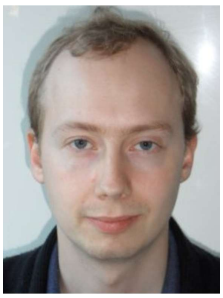
Open Papers 7

AI-Driven Predictions of Mathematics and Science GCSE Exam Results using Mock Papers **7**
Dr Sebastian Nastuta 7

Student approaches to annotation in onscreen assessments: Exploring annotation techniques and their impact on test-taker strategies **9**
Zoe Mair and Irene Custodio 9

Exploring the relationship between students' use of ICT and performance in PISA Mathematics, Reading and Science digital assessments **11**
Irene Custodio, Liyuan Liu and Sebastian Nastuta 11

Presenters



Blake Ashworth
Assessment Research Analyst

Irene Custodio
Digital Assessment Design Lead

Hayley Dalton
Head of Vocational and Assessment Research

Liyuan Liu
Senior Assessment Researcher

Zoe Mair
Digital Product Manager

Kevin Mason
Senior Assessment Researcher

Sebastian Nastuta
Data Scientist

Posters

Test Equating for maintaining standards over time in similar qualifications using Comparative Judgement

Kevin Mason

In English schools, there are often several options available to students aiming to achieve the same qualification. Standards are maintained using statistical indicators. This poster describes an approach where additional evidence is gathered using comparative judgement to ensure grades awarded across two versions of a GCSE English qualification represent the same standard.

We consider two distinct versions of GCSE English, both meeting the same content and objectives. “English 2.0” engages students with contemporary texts, attracting older students and those re-sitting. Supplementary evidence is needed to support grade setting decisions. Equating the assessments of the two qualifications is essential, but traditional approaches are not applicable as there are no common questions.

In 2024, comparative judgement exercises were conducted using assessments from both qualifications. Judges compared pairs of assessments, with each piece of work part of 22 or 23 pairs. Judges with infit values greater than 1.2 were removed from the analysis. Separation reliability was around 0.96. Theta values were regressed against qualification marks. These theta estimations were bootstrapped in order to obtain a range of possible marks, in the regressions. The range found was around seven marks in the middle of the range.

Theoretical Framework for Digital-First Assessment

Dr Liyuan Liu, Irene Custodio

As awarding organisations in the UK look to transform from paper-based and dual-mode delivery to digital-first assessments, questions are posed around how we evaluate their validity. Inspired by Bandalos's (2018) book, this poster describes a framework that seeks to integrate best practices from international large-scale assessments into the national context, emphasising reliability, validity, and accessibility as essential elements during the digital transition.

This poster introduces a structured six-step methodological tool designed to collect comprehensive research evidence for developing dual modes and digital-first assessments that are psychometrically sound and accessible, including: (1) Assessment Design and Construct Validity; (2) Content Validity and Consequences of Testing; (3) Usability and Accessibility; (4) Reliability Analysis; (5) Validity Argument Framework; and (6) Iterative Improvement. The poster clearly defines and details the evidence required for the initial four steps, subsequently illustrating how step five synthesises these findings into a cohesive validity argument. Finally, the sixth step emphasizes iterations, informed by research findings and user feedback, to ensure ongoing enhancement of assessment quality.

Overall, the methodological tool provides a systematic and rigorous approach to designing and developing digital-first assessments that uphold high psychometric standards and accessibility.

Using longitudinal datasets to explore labour market outcomes for learners taking Vocational and Technical Qualifications in England

Blake Ashworth, Hayley Dalton, Ria Bhatta

This poster reports on research looking at employment earnings for learners completing upper secondary vocational and technical qualifications (VTQs) in England. Using the Longitudinal Education Outcomes datasets (LEO), we can see where VTQs have the most impact in respect of median salary for learners going directly to work and for those accessing university before entering the labour market.

In the past decade, several studies, e.g. The Nuffield Foundation (2022) have shown the impact of VTQs on supporting progression to Higher Education. Much less is reported about longer-term employment outcomes for learners that take VTQs in respect of salary outcomes and impact on employment pipelines in skill-shortage areas such as health and digital sectors.

Regression and multi-level modelling, along with descriptive statistics have been used to look at salary outcomes through interactions that are known to impact salary, these include: economic disadvantage, lower secondary attainment, gender, SEND status, and sector/subject studied. We also spotlight the employment sectors where learners taking these qualifications and progressing to university have added the most and least salary benefits to learners, intersecting these with learner characteristics where the matched data allows.

AI-Driven Predictions of Mathematics and Science GCSE Exam Results using Mock Papers

Dr Sebastian Nastuta

Background

“Machine learning, sometimes called statistical learning, is a sub-field of artificial intelligence whereby algorithms learn patterns in data to perform specific tasks” (Rhys, 2020, 25). This research project is meant to contribute to the growing body of literature in which machine learning and artificial intelligence are used in the educational context, in this case, to analyse previous outcomes and predict future exam results. Before gaining international prominence in 2022, AI was considered “the new kid on the block” in educational context which was surprising, according to Richardson and Clesham “because there is obvious value in harnessing contemporary computing power to facilitate AI and automated machine decision-making in the processing of data such as exam results” (2021, 2). Although Artificial Intelligence is nowadays extensively used in educational contexts, mainly for content generation purposes, its power to predict students' future performance remains something beyond the reach of teachers and educational institutions.

Findings

This paper proposes a novel approach to predicting student performance in GCSE Mathematics and Science exams using AI and machine learning. We evaluate the extent to which and the accuracy with which future exam results can be predicted using mock results, a common practice in the UK. This research has the potential to significantly impact educational institutions, providing them with valuable assistance for high-stakes exam preparation.

Using a supervised Machine Learning approach, we used item-level performance from GCSE Mathematics and Science June 2023 to predict outcomes in June 2024 for students who took a shadow paper1 of the 2023 papers as mock exams. A similar approach was employed previously with 2022 data to make predictions for 2023, and using 2024 exam data, we will be making predictions for summer 2025 students.

Although using mock results marked by teachers is a common practice in the UK, this analysis relies on a set of mock papers marked by professional examiners using the same marking standards as in a real exam through our company’s dedicated Mocks Service. Additionally, we matched the predicted results with the actual outcomes to observe how

students' results evolved from formative mock results to summative high-stakes results. This potentially allows not only to make accurate predictions but also to gather information about students' knowledge gaps and areas to be improved before exams.

Comparing the predicted grades with the actual grades for at least 1700 matched candidates in four different exams, we concluded that different machine learning algorithms (OLS regression, logistic regression, support vector machine, decision tree, K-Nearest Neighbors (KNN), etc.) can provide 80% to 95% accuracy predictions. Compared with a study using a similar framework in which the proposed model achieved a classification accuracy of 70–75% (Yagci, 2022), our model performs exceptionally well, indicating a strong potential for future development.

Footnote: 1 A "shadow paper" represents a replica of a previous exam paper in which items' marks distribution, paper structure and assessment objectives are extremely similar with the original paper.

Reference

Gardner, J., O'Leary, M., Yuan, L. (2021) Artificial intelligence in educational assessment:

'Breakthrough? Or buncombe and ballyhoo?'. Journal of Computer Assisted Learning, 2021; 37: 1207–1216.

Richardson, M., Clesham, R. (2021) 'Rise of the machines?' The evolving role of AI technologies in high-stakes assessment. London Review of Education, 19 (1), 9, 1–13.

Rhys, H. (2020). Machine Learning with R, the tidyverse, and mlr, Manning Publications.

Yagci, M. (2022) Educational data mining: prediction of students' academic performance using machine learning algorithms, Smart Learning Environments (2022) 9: 11.

Student approaches to annotation in onscreen assessments: Exploring annotation techniques and their impact on test-taker strategies

Zoe Mair and Irene Custodio

As part of a transition to more digital forms of high-stakes assessment in the current educational landscape, it is important to understand how test-taker strategies may differ when interacting with both digital and paper-based formats. While exploring students' assessment outcomes and attainment across onscreen and paper-based formats, it is important to develop an in-depth understanding of the end-to-end process that students engage with to answer high-stakes exam questions, such as accessing, interpreting and interacting with question material [1]. One such activity which may be employed by students when interacting with test material is the annotation of texts.

Annotation is characterised by an active engagement with the text, and can take many forms, including underlining or highlighting words or phrases and making notations on or around text. In paper-based exams, annotation has been found to increase student confidence in their critical reading skills and positively impact comprehension and assessment performance [2]. When considering how this applies to onscreen exams, existing research has revealed differences in both the visual output of onscreen annotations and user motivation for utilising onscreen annotations compared to paper-based mark-up. However, much of this existing literature tends to focus on Maths or Science subjects [3]; given the differences in assessment structure and response formats between STEM and literary domains, these findings may not generalise to English Language assessments.

With this in mind, our research explores the potential differences between student annotation on paper-based and onscreen versions of long-form texts in the context of International GCSE English examinations. Our research also aims to explore whether classroom teaching strategies can influence students' approaches to annotation. Finally – if there are key differences in annotation strategies across different formats – do these differences impact students' engagement with test materials?

To answer these questions, we embarked on a multi-phase programme of research across two scholastic years.

Phase one compared annotation strategies demonstrated by students sitting International GCSE English Literature and International GCSE English Language exams (School A: n = 74; School B: n = 79) across two secondary schools. School A completed the exam onscreen but with a paper copy of the source texts, while School B completed the exam using only the onscreen platform. Students' annotations on both the paper-based and the onscreen

texts were collected directly following the exam. These were subsequently reviewed, and all types of annotations and markings were categorised to understand how they differed across the two formats.

Preliminary findings indicated that the overall percentage of students who opted to use annotations within their exam was consistent across both formats. A high-level examination of the annotations revealed that students employed strategies which were, as expected, visually different across the modes, yet with seemingly similar motivations e.g. marking out the referenced passages in the source text which were relevant to each question. Interestingly, across both assessment formats, students were more likely to annotate longer-form as opposed to shorter texts. Taking these findings together, Phase two was refined to further understand the different strategies employed by students when annotating long texts on paper and onscreen during assessments.

In phase two, students completed onscreen International GCSE English questions associated with long non-fiction texts of around 1000 words each. Observations and post-test, in-depth semi-structured interviews were carried out to explore students' preferences and motivations for annotations, providing rich insights into their cognitive processes and test-taker strategies.

In the final phase of the research, we examine whether the observations and findings from phase one and phase two are due to fundamental differences between onscreen and paper-based assessment and the tools available to students, or whether differences reflect students' level of familiarity with and use of onscreen tools. We also explore whether these differences can be bridged by effective demonstration and teaching of onscreen annotation methods in the classroom environment.

In this presentation, we compare student experience, test-taker strategies and student motivations when interacting with texts onscreen or on paper. In doing so, we suggest important considerations about the fairness and validity of assessments across different formats.

References

[1] Pollitt, A. and Ahmed, A. (1999) 'A New Model of the Question Answering Process'. A paper presented to the International Association for Educational Assessment, May 1999.

[2] Bedford, A. and Park-Ballay, C. (2023) 'Annotation for critical reading: An action research project.' In Australian Journal of English Education, 58(2), pp.69-80.

[3] Threlfall, J., Pool, P., Homer, M. and Swinnerton, B. (2007) 'Implicit aspects of paper and pencil mathematics assessment that come to light through the use of the computer'. Educational Studies in Mathematics, 66, pp.335-348.

Exploring the relationship between students' use of ICT and performance in PISA Mathematics, Reading and Science digital assessments

Irene Custodio, Liyuan Liu and Sebastian Nastuta

Given the increasing use of digital technologies both inside and outside of the classroom, the relationship between digital skills and achievement in digital assessment continues to be of significant interest to educators and policymakers. As education communities increasingly adopt digital forms of assessment, it is important to understand the relationship between performance in these assessments and students' perceptions, usage, and interactions with ICT (Information and Communication Technology). Previous research has yielded mixed findings regarding the impact of ICT use on students' academic performance, suggesting that this complex relationship may be influenced by various factors including the purposes and quality of ICT use, as well as students' interests, attitudes, confidence, and competencies. For example, increased availability and use of ICT, both in and outside of school, were found to negatively associate with learning outcomes. Whereas students' positive attitudes towards ICT were found to have a strong positive relationship with their academic achievements [1].

A previous study, using PISA (Programme for International Student Assessment) 2022 Mathematics data for England, included items that went beyond 'paper behind glass' designs, requiring higher interactivity and greater digital skills through visual representations, equation editors, and multi-step calculations. The findings showed that the ways in which students use digital technology matters more than availability itself. General ICT availability and usage were found to negatively impact Mathematics performance, while subject-specific ICT use and enquiry-based learning activities had a positive effect [2].

In our research, we explore the relationship between different ICT factors and student performance in PISA digital Mathematics, Science and Reading assessments. Considering that different types of digital skills may be relied upon when completing digital items across the three domains of Mathematics, Reading and Science, we seek to understand if ICT factors differentially impact performance across the three PISA subject domains. We draw upon the rich data collected in the PISA 2022 questionnaires, particularly the ICT questionnaire, devised for PISA 2022. This ICT questionnaire provides comprehensive self-reported measures of different aspects, including: ICT availability; students' attitudes towards ICT; use of ICT for learning and for leisure in and out of school; views on the ICT environment at home and at school; and students' self-efficacy when using ICT to perform various tasks [3]. The large and representative nature of the PISA 2022 dataset allows us to

build composite constructs to explore the relationship between students' self-reported measures and their achievement in digital Mathematics, Reading and Science assessments.

This research included carrying out analyses in line with the PISA Technical Report [5] to develop 12 ICT composite constructs, using the means of a series of questionnaire items. These constructs represent various aspects of ICT access, usage at school and home, as well as students' interest, self-efficacy, and competencies in ICT. We then used descriptive analyses to assess students' engagement levels with different types of ICT, followed by correlation analyses to measure the relationships between these ICT constructs and students' performance in the PISA digital assessments in each of the three subject domains. Considering the nested nature of the PISA data, as a third step, we employed two-level multilevel analysis (MLA) to examine relations between students' performance and variables measured at both student (e.g., gender, ethnicity, socioeconomic status) and school levels (e.g., the socioeconomic background of the school). Using multilevel modelling allows for the observation of variances and the simultaneous estimation of impacts across different levels [4].

In this session, we present our research findings on the relationship between students' ICT access, use, and perceived ICT competencies and performance in PISA 2022 digital assessments for Mathematics, Science, and Reading. We highlight how quality access to ICT resources and student self-efficacy positively impact performance, while frequent and excessive ICT use at home and at school negatively affects performance across the three subject domains. By examining students' use and perceptions of their ICT competencies, we shed light on the factors that impact students' performance in digital assessments. We explore implications for digital assessment design and assumptions about students' digital proficiency and their educational outcomes.

References

[1] Courtney, M., Karakus, M., Ersozlu, Z. and Nurumov, K., 2022. The influence of ICT use and related attitudes on students' math and science performance: multilevel analyses of the last decade's PISA surveys. Large-scale Assessments in Education, 10(1), p.8.

[2] Custodio, I., Grima, G., Liu, L., and Nastuta, S., 2024. Exploring the relationship between students' use of digital technologies and their performance in digital PISA 2022 mathematics assessments. 2024 AEA-Europe Annual Conference.

[3] Lorenceau, A., C. Marec and T. Mostafa (2019), "Upgrading the ICT questionnaire items in PISA 2021", OECD Education Working Papers, No. 202, OECD Publishing, Paris, https://doi.org/10.1787/d0f94dc7-en.

[4] Muijs, D., 2010. Doing quantitative research in education with SPSS. Sage.

[5] OECD (2024), PISA 2022 Technical Report, PISA, OECD Publishing, Paris, https://doi.org/10.1787/01820d6d-en.