CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Research Matters

Issue 37 / Spring 2024

Proud to be part of the University of Cambridge

Cambridge University Press & Assessment unlocks the potential of millions of people worldwide. Our qualifications, assessments, academic publications and original research spread knowledge, spark enquiry and aid understanding.

Citation

Articles in this publication should be cited using the following example for article 1: Li, X. (2024). Extended Reality (XR) in mathematics assessment: A pedagogical vision. *Research Matters: A Cambridge University Press & Assessment publication, 37*, 6–23. https://doi.org/10.17863/CAM.106031

Credits

Reviewers: Matthew Carroll, Gill Elliott, Dominika Majewska, Tim Oates and Carmen Vidal Rodeiro

Editorial and production management: Lisa Bowett

Additional proofreading: Alison French

Cambridge University Press & Assessment is committed to making our documents accessible in accordance with the WCAG 2.1 Standard. We're always looking to improve the accessibility of our documents. If you find any problems or you think we're not meeting accessibility requirements, please contact our team: Research Division, ResearchDivision@cambridge.org

If you need this document in a different format contact us, telling us your name, email address and requirements and we will respond within 15 working days.

Research Matters Issue 37, https://doi.org/10.17863/CAM.106030

All details are correct at the time of publication in March 2024.

Contents

Foreword

Tim Oates, CBE

The explosion of activity around large language models since the release of ChatGPT in November 2022 has pushed some very important assessment issues aside. It is vital not to forget them. Issues such as the “speededness” of examinations - explored in this edition of Research Matters remain fundamental to both measurement and the quality of the experience of candidates during assessment. It's a reminder that core principles of assessment - around validity, reliability, utility, bias and fairness - apply even in times of rapid and exciting development. Digitally-supported and enhanced assessment are likely to supply important, long-anticipated innovations in areas such as adaptive assessment and the “seamless integration of assessment into learning...” But as innovation increases from outside of established communities of assessment researchers and developers, it is vital that the carefully accumulated scientific understanding of the quality of assessment does not take a backseat to impressive technical innovation. The reason for this is simple: assessment relies on trust - at so many different levels: from the candidate expecting high quality, accessible questions and tasks to the selector assuming the dependability of the information and signals from a qualification report, score or grade. The principles and criteria for high quality assessment - whether low or high stakes in character - have been carefully accumulated through experience and research. They are central to the international portability of qualifications. Breakdown of trust in assessments would compromise so many of the domestic and international functions which we have so carefully accumulated: fair recognition of attainment, unlocking of progression, and efficient supply of skills and knowledge to economies. It's extremely important to remember that innovation in assessment should rest on solid foundations.

Editorial

Tom Bramley

Welcome to the spring issue of *Research Matters*. Much current debate in education and assessment is around the potential of technology to enhance (or otherwise) student learning. Our first two articles are on this theme. The first, by Xinyue Li from Cambridge Mathematics, describes the technologies collectively known as "extended reality" and considers opportunities and challenges for using them in teaching and assessing mathematics. The second, by Jude Brady and colleagues, reports on a study where three undergraduates were asked to use ChatGPT to assist with writing essays and then interviewed about their approach.

Our third article, by Nicky Rushton, Dominika Majewska and Stuart Shaw, considers the difficult issues that arise when comparing curriculum documents with the aim of making claims about comparability of different curricula. In particular, they focus on the application of the "mapping method", using a comparison between the Common Core State Standards in the US and the mathematics national curriculum in England as an example.

Computer-based testing affords the possibility of collecting evidence not only of the student's response itself, but of other features of the process that produced the response, such as the time taken for each question. With paper-based examinations, however, we usually do not know how long it took students to complete their answers. In our fourth article Emma Walland explores the extent to which data (specifically whether a response was missing or not) can support inferences about whether students were under time pressure in paper-based GCSE examination components, and whether exams in some subjects were more "speeded" than others.

Our final article, by Chris Jellis, presents a historical overview of the Centre for Evaluation and Monitoring (CEM), acquired by Cambridge in 2019 but now celebrating more than 40 years since its creation. It provides a fascinating insight into CEM's role in pioneering ways for schools to evaluate their effectiveness, and its contribution to some key assessment debates over the years.

Extended Reality (XR) in mathematics assessment: A pedagogical vision

Xinyue Li (Cambridge Mathematics)

The year of 2023 saw a surge of interest in artificial intelligence (AI), with "Hallucinate" being named as Word of the Year 2023 by *Cambridge Dictionary* (Cambridge University Press & Assessment, n.d.), reflecting a growing curiosity about emerging technologies that have the potential to significantly alter human perception and experience (Li & Zaki, 2024). In particular, as stated in the *Futures of Assessment* report, advancements in technology are transforming assessment methods and the vision is that learners in 2050 may be immersed in an educational environment where augmented, virtual and hybrid technologies are comprehensively embedded in assessments (Abu Sitta et al., 2023). Against this background, extended reality (XR) – encompassing virtual reality (VR), augmented reality (AR), and mixed reality (MR) – emerges as a potential transformative tool in educational realms. This article explores the potential of XR in facilitating mathematics assessments; it proposes a list of mathematical topics that could be effectively mediated by XR's immersive and interactive features. Additionally, it discusses some major challenges which could be barriers to the widespread adoption of XR in educational contexts and sets out a research agenda for further investigation.

Definition of XR

Extended reality (XR) is "an emerging umbrella term for all the immersive technologies" (Marr, 2019). As the landscape of technological innovation continually evolves, defining XR remains a moving target (Palmas & Klinker, 2020). Currently, XR refers to established technologies including AR, VR, and MR (Lee, 2020), as well as those yet to be developed.

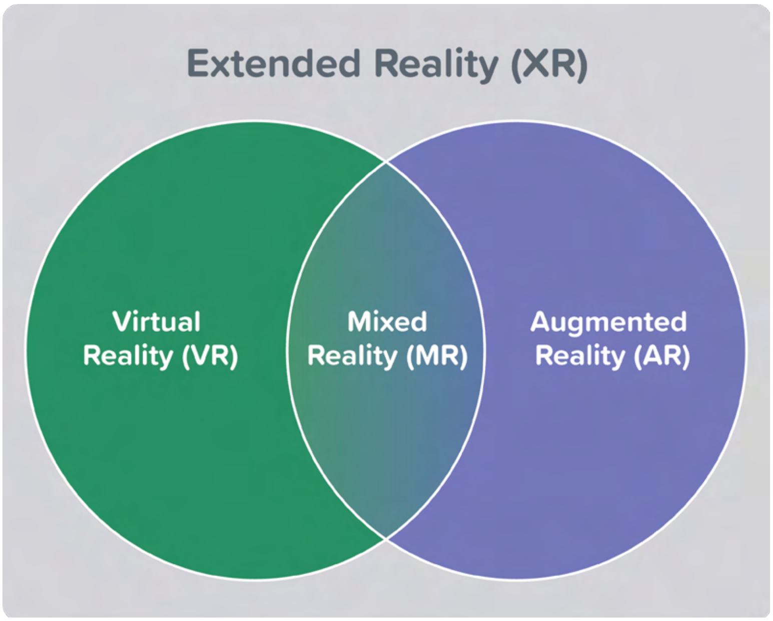


Figure 1: How VR, AR and MR intersect

Types and core features of XR

Virtual Reality

In the *Cambridge Academic Content Dictionary* (Cambridge University Press, 2017), virtual reality (VR) is defined as "a set of images and sounds produced by a computer that seem to represent a real place or situation"; therefore, VR "provides a computer-generated environment wherein the user can enter a virtual environment with a VR headset and interact with it" (Rokhsaritalemi et al., 2020, p. 1).

Augmented Reality

Augmented reality (AR) is a technology that enables the real-time integration of computer-generated virtual elements with either a direct or indirect view of the real world (Azuma, 1997; Lee, 2012). AR-based content can span multiple sensory modalities; for example, visual, auditory, haptic, etc. (Cipresso et al., 2018). While the lack of relation to real space is one of the characteristics of VR, AR presents a new method of visualisation that allows for the addition of computer-generated content to the real world (Rokhsaritalemi et al., 2020, p. 1).

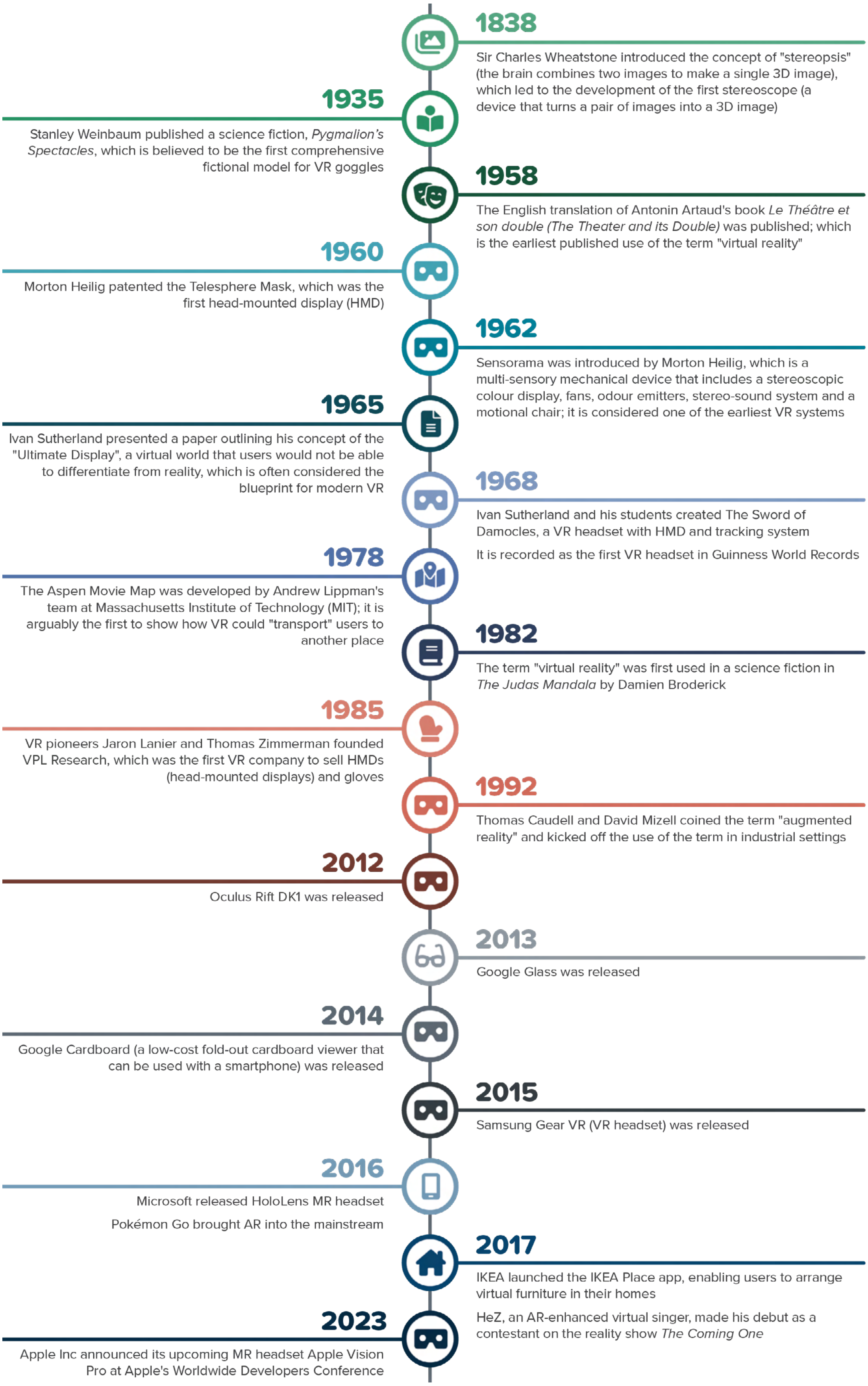
Mixed Reality

The term mixed reality (MR) was introduced by Paul Milgram and Fumio Kishino in their paper "A Taxonomy of Mixed Reality Visual Displays" (1994). It can be understood as a blend of physical and digital worlds, which is based on "advancements in computer vision, graphical processing, display technologies, input systems, and cloud computing" (Microsoft, 2023).

Table 1 below summarises the core features of VR, AR, and MR.

Table 1: Comparison of VR, AR and MR (Developed from Jaquith, 2016; Li & Taber, 2022; Li & Zaki, 2024; McMillan et al., 2017; Rokhsaritalemi et al., 2020; Taber & Li, 2021.)

Features	Virtual Reality (VR)	Augmented Reality (AR)	Mixed Reality (MR)
Display device	Special headsets or glasses required in most situations	Special headsets are optional; can be viewed through a digital device (e.g., a smartphone, a tablet, etc.)	Special headsets are optional
Image source	Computer-generated graphics	Combination of computer-generated elements and real-life elements	Combination of computer-generated elements and real-life elements
Environment	Fully digital/virtual	Real-life and virtual elements are blended seamlessly	Real-life and virtual elements are blended seamlessly
Perspective	Virtual elements will change their position and size according to the user's perspective in the digital/virtual world	Virtual elements are experienced based on the user's perspective in the real world	Virtual elements are experienced based on the user's perspective in the real world
Presence	Feeling of being "transported" to a different location with no sense of the real world the user is in	The user remains aware of the real world, with virtual elements added to their view	The user feels present in the real world with superimposed virtual elements
Awareness	The user cannot see elements of the real world while immersed in VR	The user can identify virtual elements based on their nature and behaviour (e.g., floating text)	The user interacts with virtual elements as if they are part of the real world



1838

Sir Charles Wheatstone introduced the concept of "stereopsis" (the brain combines two images to make a single 3D image), which led to the development of the first stereoscope (a device that turns a pair of images into a 3D image)

1935

Stanley Weinbaum published a science fiction, *Pygmalion's Spectacles*, which is believed to be the first comprehensive fictional model for VR goggles

1958

The English translation of Antonin Artaud's book *Le Théâtre et son double (The Theater and its Double)* was published; which is the earliest published use of the term "virtual reality"

1960

Morton Heilig patented the Telesphere Mask, which was the first head-mounted display (HMD)

1962

Sensorama was introduced by Morton Heilig, which is a multi-sensory mechanical device that includes a stereoscopic colour display, fans, odour emitters, stereo-sound system and a motional chair; it is considered one of the earliest VR systems

1965

Ivan Sutherland presented a paper outlining his concept of the "Ultimate Display", a virtual world that users would not be able to differentiate from reality, which is often considered the blueprint for modern VR

1968

Ivan Sutherland and his students created The Sword of Damocles, a VR headset with HMD and tracking system

It is recorded as the first VR headset in Guinness World Records

1978

The Aspen Movie Map was developed by Andrew Lippman's team at Massachusetts Institute of Technology (MIT); it is arguably the first to show how VR could "transport" users to another place

1982

The term "virtual reality" was first used in a science fiction in *The Judas Mandala* by Damien Broderick

1985

VR pioneers Jaron Lanier and Thomas Zimmerman founded VPL Research, which was the first VR company to sell HMDs (head-mounted displays) and gloves

1992

Thomas Caudell and David Mizell coined the term "augmented reality" and kicked off the use of the term in industrial settings

2012

Oculus Rift DK1 was released

2013

Google Glass was released

2014

Google Cardboard (a low-cost fold-out cardboard viewer that can be used with a smartphone) was released

2015

Samsung Gear VR (VR headset) was released

2016

Microsoft released HoloLens MR headset

Pokémon Go brought AR into the mainstream

2017

IKEA launched the IKEA Place app, enabling users to arrange virtual furniture in their homes

HeZ, an AR-enhanced virtual singer, made his debut as a contestant on the reality show *The Coming One*

2023

Apple Inc announced its upcoming MR headset Apple Vision Pro at Apple's Worldwide Developers Conference

Figure 2: A brief history and evolution of XR

The research landscape

To establish a broad understanding of the research landscape of XR use in mathematics assessment, a systematic literature review was conducted on the *Web of Science* database using “extended reality” or “XR”, and “mathematics assessment” as Topic (((TS=(extended reality)) OR TS=(XR)) AND TS=(mathematics assessment)), which returned no results. Therefore, to broaden the search parameters, “extended reality” or “XR”, and “mathematics education” were searched (((TS=(extended reality)) OR TS=(XR)) AND TS=(mathematics education)), yielding 27 results. All publications were reviewed for relevance to this article. Two publications used the term XR for a different purpose, five publications used the words “reality” and “extended” in contexts different from those of the present article, two publications were not closely relevant to mathematics education, and one publication focused solely on mathematics but not on XR. Consequently, these publications were excluded from Table 2. The order in Table 2 reflects the sequence shown on the *Web of Science* database.

Table 2: Summary of the literature review

Title: Adoption of virtual and augmented reality for mathematics education: A scoping review **Author(s) and publication year:** Lai, J. W., & Cheong, K. H. (2022) **Study type:** Literature review **Topic:** Implications of immersive XR on mathematics pedagogy in higher education. **Key finding:** The development of an enhanced framework for XR learning environments.
Title: XR maths – designing a collaborative extended realities lab for teaching mathematics **Author(s) and publication year:** Gilardi, M., Hainey, T., Bakhshi, A., Rodriguez, C., & Walker, A. (2021) **Study type:** Empirical study **Topic:** The design of XR applications for educational purposes (in higher education contexts). **Key finding:** A process for designing an XR application for educational purposes.
Title: Exploring the impact of extended reality (XR) on spatial reasoning of elementary students **Author(s) and publication year:** Baumgartner, E., Ferdig, R. E., & Gandolfi, E. (2022) **Study type:** Empirical study **Topic:** An investigation into the impact of XR video content on elementary students’ spatial reasoning skills. **Key finding:** The consumption and production of XR videos could improve the spatial reasoning abilities of elementary students.
Title: Coordi: A virtual reality application for reasoning about mathematics in three dimensions **Author(s) and publication year:** Pearl, H., Swanson, H., & Horn, M. (2019) **Study type:** Empirical study **Topic:** Evaluation and refinement of a VR application designed for assisting high school students in plotting points, drawing and manipulating graphs, vectors, objects, and reasoning in 3D space. **Key finding:** This VR application could enhance mathematics learning outcomes.

Title: Playable experiences through technologies: Opportunities and challenges for teaching simulation learning and extended reality solution creation
Author(s) and publication year: See, Z. S., Ledger, S., Goodman, L., Matthews, B., Jones, D., Fealy, S., Har Ooi, W., & Amin, M. (2023)
Study type: Empirical study
Topic: Simulation learning and XR solution creation skills for tertiary education students.
Key finding: Key criteria and a flexible outline for academic researchers and learning designers in higher education, focusing on XR in teaching and inclusive learning design.

Title: XRLabs: Extended reality interactive laboratories
Author(s) and publication year: Kiourt, C., Kalles, D., Lalos, A. S., Papastamatiou, N., Silitziris, P., Paxinou, E., Theodoropoulou, H., Zafeiropoulos, V., Papadopoulos, A., & Pavlidis, G. (2020)
Study type: An introduction to the XRLabs
Topic: An introduction to the XRLabs platform: an XR platform designed to aid in the training of students at all educational levels.
Key finding: The highly interactive platform enables students to engage in sustainable edutainment experiences, particularly beneficial in distance or online learning contexts for Science, Technology, Engineering, and Mathematics (STEM).

Title: Augmented reality in mathematics education: The case of GeoGebra AR
Author(s) and publication year: Tomaschko, M., & Hohenwarter, M. (2019)
Study type: Empirical study
Topic: An exploration of the potential of AR in learning and teaching mathematics, with a special emphasis on GeoGebra AR.
Key finding: Suggestions for potential future developments of the GeoGebra AR app.

Title: Pre-service teachers' professional noticing when viewing standard and holographic recordings of children's mathematics
Author(s) and publication year: Kosko, K. W. (2022)
Study type: Empirical study
Topic: An exploration of the use of holographic representations.
Key finding: Viewing holograms prior to standard videos is more beneficial than viewing standard videos first.

Title: From STEM to STEAM: An enactive and ecological continuum
Author(s) and publication year: Videla, R., Aguayo, C., & Veloz, T. (2021)
Study type: Literature review; secondary analysis on existing empirical studies
Topic: The integration of Science, Technology, Engineering, Arts, and Mathematics (STEAM) education.
Key finding: The development of an enactive and ecological approach.

Title: Kinesthetic learning applied to mathematics using Kinect
Author(s) and publication year: Ayala, N. A. R., Mendívil, E. G., Salinas, P., & Rios, H. (2013)
Study type: Empirical study
Topic: The impact of kinaesthetic learning on mathematics education.
Key finding: AR could boost the learning curve, although its effectiveness is limited by certain factors (e.g., dependency on markers, the range of movement).

Title: Comparative study of technological and communication means to improve the articulation between the secondary and university levels

Author(s) and publication year: Gómez, M. M., Saldis, N. E., Bielewicz, A., Colasanto, C. M., & Carreño, C. T. (2019)

Study type: Empirical study

Topic: An investigation into the use of computer technology and networks among high school students, particularly their perception of these tools as instruments for formal learning in mathematics.

An exploration of the development and application of various didactic materials incorporating technology to foster autonomous learning.

Key finding: The introduction of videos, guides featuring XR (QR codes), and a virtual classroom can enhance students' autonomy in learning. Among the tools tested, videos and XR were preferred, while the virtual classroom was less favoured but still effective.

Title: Using the PerFECt framework to establish an onlife community for theatre in mathematics to teach principles of computing

Author(s) and publication year: Moumoutzis, N., Paneva-Marinova, D., Xanthaki, C., Arapi, P., Pappas, N., & Christodoulakis, S. (2020)

Study type: Description of the PerFECt framework

Topic: An investigation into how modern digital platforms and applications embody new qualities and affordances, and how they can be designed to provide new capabilities to users.

Key finding: Specific design principles, with a practical example of these principles applied in the design of a community of practice for teachers.

Title: Enhancing STEM education using augmented reality and machine learning

Author(s) and publication year: Ang, I. J. X., & Lim, K. H. (2019)

Study type: Applied research

Topic: The transition of STEM education from traditional textbooks to interactive platforms utilising electronic devices (e.g., AR).

Key finding: The demonstration of how AR can be integrated into educational platforms to increase learning motivation and students' understanding of STEM subjects.

Title: Multimodal technologies in precision education: Providing new opportunities or adding more challenges?

Author(s) and publication year: Qushem, U. B., Christopoulos, A., Oyelere, S. S., Ogata, H., & Laakso, M.-J. (2021)

Study type: Literature review

Topic: An examination of the role of multimodal technologies in Personalised or Precision Education (PE).

Key finding: PE techniques could enhance the effectiveness of educational platforms and tools, facilitating the acquisition of knowledge and development of skills for students.

Title: Multimodal analysis of interaction data from embodied education technologies

Author(s) and publication year: Walkington, C., Nathan, M. J., Huang, W., Hunnicutt, J., & Washington, J. (2023)

Study type: Empirical study

Topic: The discussion of the potential of immersive digital technologies such as shared augmented reality (shAR), VR, and motion capture (MC) in enhancing the understanding of human cognition and creating innovative technology-enhanced learning experiences.

Key finding: The exploration of a multimodal analysis method for studying embodied technologies in educational technology research.

Title: Exploration of kinesthetic gaming for enhancing elementary math education using culturally responsive teaching methodologies
Author(s) and publication year: Barmpoutis, A., Ding, Q., Anthony, L., Eugene, W., & Suvajdzic, M. (2016)
Study type: Empirical study
Topic: An exploration of a novel computer-assisted culturally responsive teaching (CRT) framework specifically designed for teaching mathematics to 5th grade students.
Key finding: The development and implementation of a CRT framework that blends traditional CRT methods with modern digital technologies.

Title: Harnessing early spatial learning using technological and traditional tools at home
Author(s) and publication year: Lee, J., Ho, A., & Wood, E. (2018)
Study type: Literature review; evaluation of educational software programs
Topic: An investigation into the role of parents and early childhood educators in developing foundational mathematical concepts in young children, specifically geometry and spatial sense.
Key finding: Highlighting the importance of manipulative play in fostering creative and educational experiences for young learners.

This literature review provides insights into current practices and identifies potential gaps for future research, particularly highlighting that the use of XR in mathematics education, especially in mathematics assessment, is an under-researched field. Although all reviewed publications reported positive findings, ranging from enhanced learning motivation to effective learning outcomes when teaching and learning with XR, much of the existing literature on XR-assisted mathematics education focuses more on XR's technical aspects than on pedagogical perspectives. By reviewing the existing literature, this section plays a crucial role in setting the stage for future empirical studies that are essential to unlock the full potential of XR as a tool for facilitating effective and innovative mathematics assessment. Therefore, this literature review is not just a brief summary of current practices, but a call to action for researchers to embark on rigorous empirical studies that will provide more evidence to guide the effective integration of XR in mathematics education.

The theoretical framework: theorising XR as tools

The theorisation of technology is often missing from the canon of research in the field of technology-assisted education (Oliver, 2013), hence the need to address the topic in this article. Oliver found that there was a limited number of publications focusing on the study of technology from a theoretical perspective, and most of these attempts had drawn on the concept of affordance. Coined by James Gibson (1979), "affordance" was initially developed in the field of ecological psychology as Gibson argued that "affordances of the environment are what it offers the animal, what it provides or furnishes, either for good or ill" (p. 127).

Affordance can be understood as clues (which can be explicit/obvious or implicit/hidden) that give users hints about how to interact with certain objects. Oliver (2011), among others, argues that affordance-based accounts have positioned technology as the cause of changes in learning, which is being technologically deterministic – a concept posited by Thorstein Veblen, who believed that technology was the agent of social change. However, acknowledging other influential elements in societal growth is crucial, as it would be simplistic to attribute such significant influence solely to technology. Consequently, there is a need for an alternative account to better understand the use of digital technology in education. One of the critical responses to the beliefs that position technology as a determinant of practice is to theorise technology from social perspectives (Oliver, 2013). This is based on constructivist accounts (Thorpe, 2002), and values the agency of learners, which is absent in the deterministic perspective.

It is argued that Vygotsky's ideas are relevant to the uptake of digital technologies in learning (Taber & Li, 2021). For example, for Vygotsky, tools play the "mediating role in human reaction and interaction with the world" (Verenikina, 2010, p. 19). Tools can be categorised as external/physical tools (e.g., artefacts, instruments, etc.) and internal/psychological/symbolic tools (e.g., procedures, methods, concepts, etc.). External tools are designed to "manipulate physical objects", and internal tools can be used by learners to "influence people or themselves" (Verenikina, 2010, p. 19). For the purpose of this article, XR technologies are theorised as external tools.

According to Vygotsky, the use of tools and the ability to improve tools are important for our development as humans, and we can use tools to mediate activities (Taber, 2020). In the context of mathematics education, using a tool to mediate an activity refers to employing a specific device, software, or method to facilitate understanding, engagement, or skill development. Imagine a mathematics class focused on 3D geometry, where concepts such as the properties of 3D shapes, volume, and surface area could be abstract and challenging to understand through traditional two-dimensional (2D) textbooks. In this scenario, the use of VR headsets would enable students to "enter" a virtual space, where they can interact directly with 3D geometric shapes. This experience allows them to view, manipulate, and explore these shapes in ways not possible with a 2D textbook. Consequently, the VR tool not only facilitates a better understanding of geometry through immersive visualisation but also enhances the learning process, making it more effective and enjoyable for students.

As argued by Taber (2020), mediation plays an important role in scaffolding processes that would otherwise be unachievable. If we theorise XR technologies as tools within this context, it leads to a fundamental design principle in digital assessment in mathematics. I suggest that XR technologies should only be adopted when other digital or traditional methods are inadequate. For instance, while XR technologies could offer innovative ways to assess certain mathematics topics (see the following section for detailed examples), they might not be the most effective means for assessing all topics. Other methods, such as the paper-and-pencil approach, might be more suitable for some topics (e.g., basic arithmetic operations) due to their simplicity and directness. Therefore, it is crucial to ensure that XR technologies are used as a means to an end, rather than as an end itself.

The potential of XR in mathematics assessment

It is important to start examining XR technologies that already exist and to learn from the current use of these emerging tools and resources, drawing inferences from them about the potential use and impact of these resources in mathematics assessment, rather than waiting for them to be fully implemented in the classroom or exam hall. Therefore, this section presents a list of some possible topics that can be facilitated through the use of XR technologies in mathematics assessment. The implications and applications of XR technologies were mapped to each of these topics, as presented in Table 3 below. The topics are listed in alphabetical order. The list of topics and their associated implications is not exhaustive; it is intended to provide some of the examples.

Table 3: XR-based resources for facilitating topics in mathematics assessment (developed from Li & Zaki, 2024)

Mathematical topics	Practical implications for XR and integration
Algebra	XR could facilitate algebra-related items in mathematics assessment by enabling test-takers to solve interactive problems overlaid onto their real-world surroundings. This could involve test-takers physically manipulating variables and observing changes in real time, providing a more comprehensive assessment of their understanding and problem-solving skills. For instance, XR could facilitate assessment by initially allowing test-takers to manipulate virtual number lines and geometric representations of algebraic principles. As complexity increases, XR can introduce interactive environments for exploring polynomial factoring, with virtual manipulatives for rearranging terms, and eventually, immersive scenarios for applying algebra in real-world problem-solving, such as calculating trajectories in physics simulations.
Calculus	XR could provide an opportunity for test-takers to engage with and manipulate three-dimensional (3D) mathematical constructs, giving them a live opportunity to demonstrate their understanding of complex concepts such as integrals and derivatives through direct interaction with virtual models. For instance, XR resources may start with visualising the concept of limits by illustrating approaching curves and dynamically showing how values change. For derivatives, test-takers could interact with a 3D graph, physically adjusting the slope of tangents. For integrals, XR could simulate filling volumes under curves, with real-time feedback on the calculations. Assessments could involve test-takers optimising 3D printed structures by applying differential calculus to determine stress points.

Geometry	XR excels in rendering 3D shapes, allowing test-takers to explore and understand geometric properties and theorems in a more intuitive and tangible way. In addition, by interacting with geometric figures in a virtual space, test-takers can develop stronger spatial reasoning skills, crucial for understanding concepts like angles, symmetry, and transformations. This virtual hands-on approach provides a more practical assessment of their ability to understand and apply geometric theorems to both virtual and physical spaces. For instance, XR could enable test-takers to explore the properties of 3D objects by rotating, combining, and dissecting them in virtual space. In addition, it may also include solving interactive puzzles that require applying theorems or calculating areas and volumes of complex shapes overlaid onto the physical classroom.
Statistics	XR could bring a new dimension to statistics-related assessment items, offering test-takers the opportunity to engage with interactive graphs and datasets that integrate seamlessly into their real-world surroundings. This enables a practical evaluation of their ability to interpret and analyse data in an immersive context. For instance, XR could introduce concepts such as mean, median and mode through visual, interactive plots that test-takers can alter by adding or removing data points. In addition, they could use XR to design and conduct virtual experiments, visualise probability distributions, and perform regression analyses with guided tutorials (this can be mediated with artificial intelligence-based tools). Test-takers might be asked to interpret 3D graphs of statistical data projected in the classroom, explaining their insights and conclusions.
Probability theory	XR could create engaging assessment scenarios where test-takers can experiment with and predict outcomes within virtual simulations that are overlaid onto their real-world surroundings, facilitating test-takers' conceptual understanding of probability and their practical applications, and offering a comprehensive evaluation of their problem-solving skills and theoretical understanding. For instance, XR could assist in understanding probability through simple games of chance, like dice rolls and coin flips, with visual representations of outcomes. Test-takers could also engage in complex simulations such as predicting weather patterns or market trends, or risk assessment in business contexts.

Challenges and limitations

As the use of XR in mathematics assessment is currently an under-researched and under-designed field, the absence of rigorous studies limits our understanding of its potential, challenges and limitations. Therefore, this section aims to provide an overview of the challenges and limitations that XR poses in the field of education, rather than solely focusing on mathematics assessment.

Accessibility and scalability

Accessibility remains a significant challenge in implementing XR (Biswas et al., 2021). While schools might be able to supply the necessary hardware and software for test-takers during assessment conducted within the school premises, not every test-taker has access to these resources for practising or revision purposes in out-of-class contexts.

In addition, many researchers have pointed out the limitation of scalability (e.g., Scavarelli et al., 2019). As XR technologies are rapidly evolving, schools need to update the content in assessment continually to keep up with the latest advancements.

Content validity

While XR offers immersive and interactive experiences, there is a risk of overstimulation or distraction, as test-takers might focus more on the novelty of the technology rather than the mathematical topics and skills being assessed. Against this background, it is important to ensure content validity, which could be achieved if assessment items are well aligned with both the subject matter and the required cognitive skills. Therefore, it is crucial to balance the technical engagement with educational objectives when adopting XR technologies in mathematics assessment.

Cost

One of the primary barriers to the widespread adoption of XR in educational contexts is the cost (Al-Ansi et al., 2023). High-quality XR systems require a significant financial investment; the cost of developing and purchasing the necessary equipment, along with its maintenance and regular updates, can be prohibitive for many stakeholders.

Infrastructure

One of the primary challenges in implementing XR in educational contexts is the need for robust infrastructure; like all digital technologies, XR requires robust IT support to maintain and troubleshoot (Al-Ansi et al., 2023). To fully implement XR in mathematics assessment, advanced hardware and software are required to support XR-assisted assessment items. This would normally include high-performance computers, VR/AR/MR headsets or glasses, and a stable internet connectivity.

Interdisciplinary and multidisciplinary collaboration

The utilisation of XR in mathematics assessment presents the challenge of the necessity for an interdisciplinary and multidisciplinary approach in the design process. As Gilardi et al. (2021) highlighted, an effective XR design team must comprise professionals with diverse expertise, including education, graphic and interaction design, and research methods. This implies a significant investment in assembling a team with the right skill set.

Motion sickness

Due to XR technology's immersive nature, it can cause motion sickness (when a user's senses fall out of sync) for certain users (Carter, 2023). This can occur when there is a disconnect between what users see in the virtual environment and their physical perception, leading to discomfort and disorientation. This issue can hinder the learning process and may exclude some test-takers from fully participating in XR-assisted assessment. However, it is possible that this can be reduced by shortening the time of the engagement and allowing test-takers to take regular breaks between stages.

Training

Effective implementation of XR in mathematics assessment requires teachers, practitioners and educators to be adequately trained (Li & Zaki, 2024). This would include not only the technical know-how of operating XR equipment but also the pedagogical skills to integrate these technologies effectively into the curriculum and assessment. In addition, it is also crucial to ensure that the IT support staff are adequately trained to handle any arising issues. This can be a significant challenge for the widespread adoption of XR in educational contexts.

Future directions and recommendations

Based on the discussions presented in this article, this concluding section proposes a research agenda for the widespread adoption of XR in mathematics assessment. This agenda contains various dimensions of how XR technologies can support, enhance and transform mathematics assessment. Some of these dimensions and suggested research foci are briefly presented below, organised in alphabetical order.

Accessibility and inclusivity in XR assessments

- To assess the accessibility of XR technologies for students with special learning needs.
- To explore how XR technologies could be tailored to reflect diverse learning needs in mathematics assessment.

Comparative studies on XR-assisted vs traditional assessment methods

- To investigate the efficacy of XR-assisted mathematics assessment compared to traditional paper-based or other means of digital assessment (e.g., computer-based assessment, etc.).
- To examine the impact of XR technologies on test performance (e.g., reaction speed, depth of understanding, etc.).
- To explore the optimal balance between immersive experience and information processing to avoid overwhelming test-takers.

Longitudinal impact of XR on learning trajectories

- To conduct longitudinal studies to understand the long-term effects of test-takers' engagement with XR technologies on their progression in mathematics learning.
- To evaluate how continued exposure to XR technologies could influence test-takers' attitudes towards mathematics, learning motivation and their self-efficacy.

Scalability and implementation in educational settings

- To evaluate the scalability of XR technologies in schools, considering factors such as cost, infrastructure and teacher readiness.
- To investigate best practices for the adoption and integration of XR-assisted mathematics assessments at various educational levels (e.g., primary, secondary, higher education, etc.).

XR-assisted mathematics assessment design principles

- To develop and refine guidelines for creating effective XR assessment tools.
- To investigate how different design elements (e.g., interactivity, feedback mechanisms, etc.) could influence test performance.
- To understand how interaction patterns with XR can provide insights into test-takers' mathematical thinking processes.
- To foster innovation in XR content creation that aligns with mathematics curriculum standards and assessment criteria.

Acknowledgements

I am grateful to Darren Macey, Dr Rachad Zaki, Lynn Fortin, Ray Knight, and Prof. Keith Taber for their support, enlightenment, encouragement and insightful suggestions which contributed to the shaping of this work.

References

Abu Sitta, F., Maddox, B., Casebourne, I., Hughes, S., Kuvalja, M., Hannam, J., & Oates, T. (2023). *The futures of assessment: Navigating uncertainties through the lenses of anticipatory thinking*. Digital Education Futures Initiative Cambridge; Cambridge University Press & Assessment.

Al-Ansi, A. M., Jaboob, M., Garad, A., & Al-Ansi, A. (2023). Analyzing augmented reality (AR) and virtual reality (VR) recent development in education. *Social Sciences & Humanities Open, 8*(1), 100532.

Ang, I. J. X., & Lim, K. H. (2019). Enhancing STEM education using augmented reality and machine learning. In *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, 1–5, IEEE.

Ayala, N. A. R., Mendívil, E. G., Salinas, P., & Rios, H. (2013). Kinesthetic learning applied to mathematics using Kinect. *Procedia Computer Science, 25*, 131–135.

Azuma, R. T. (1997). A survey of augmented reality. *Presence: Teleoperators and Virtual Environments, 6*(4), 355–385.

Barmpoutis, A., Ding, Q., Anthony, L., Eugene, W., & Suvajdzic, M. (2016). Exploration of kinesthetic gaming for enhancing elementary math education using culturally responsive teaching methodologies. In *2016 IEEE Virtual Reality Workshop on K-12 Embodied Learning through Virtual & Augmented Reality (KELVAR)*, 1–4, IEEE.

Baumgartner, E., Ferdig, R. E., & Gandolfi, E. (2022). Exploring the impact of extended reality (XR) on spatial reasoning of elementary students. *TechTrends, 66*(5), 825–836.

Biswas, P., Orero, P., Swaminathan, M., Krishnaswamy, K., & Robinson, P. (2021). Adaptive accessible AR/VR systems. In *Extended abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, Article number 92.

Cambridge University Press & Assessment. (n.d.). Hallucinate. In Cambridge Dictionary.

Cambridge University Press. (2017). Virtual reality. In Cambridge Academic Content Dictionary.

Carter, R. (2023, July 3). XR Today's guide to stopping VR motion sickness. *XR Today*.

Cipresso, P., Giglioli, I. A. C., Raya, M. A., & Riva, G. (2018). The past, present, and future of virtual and augmented reality research: A network and cluster analysis of the literature. *Frontiers in Psychology, 9*.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin Harcourt.

Gilardi, M., Hainey, T., Bakhshi, A., Rodriguez, C., & Walker, A. (2021). XR Maths – designing a collaborative extended realities lab for teaching mathematics. In P. Fotaris (Ed.), *Proceedings of the 15th European Conference on Games Based Learning*, 277–286. Academic Conferences and Publishing International Ltd.

Gómez, M. M., Saldis, N. E., Bielewicz, A., Colasanto, C. M., & Carreño, C. T. (2019). Comparative study of technological and communication means to improve the articulation between the secondary and university levels. *Virtualidad Educacion y Ciencia, 10*(18), 100–116.

Jaquith, T. (2016, August 17). VR, AR, and MR: What's the difference? *Futurism*.

Kiourt, C., Kalles, D., Lalos, A. S., Papastamatiou, N., Silitziris, P., Paxinou, E., Theodoropoulou, H., Zafeiropoulos, V., Papadopoulos, A., & Pavlidis, G. (2020). XRLabs: Extended reality interactive laboratories. In *CSEDU*, 1, 601–608.

Kosko, K. W. (2022). Pre-service teachers' professional noticing when viewing standard and holographic recordings of children's mathematics. *International Electronic Journal of Mathematics Education, 17*(4).

Lai, J. W., & Cheong, K. H. (2022). Adoption of virtual and augmented reality for mathematics education: A scoping review. *IEEE Access, 10*, 13693–13703.

Lee, H. (2020). A conceptual model of immersive experience in extended reality. *PsyArXiv*.

Lee, J., Ho, A., & Wood, E. (2018). Harnessing early spatial learning using technological and traditional tools at home. *Creativity and technology in mathematics education*, 279–302.

Lee, K. (2012). Augmented reality in education and training. *TechTrends, 56*, 13–21.

Li, X., & Taber, K. S. (2022). The future of interaction: Augmented reality, holography and artificial intelligence in early childhood science education. In S. Papadakis & M. Kalogiannakis (Eds.), STEM, robotics, mobile apps in early childhood and primary education (pp. 415–442). Springer.

Li, X., & Zaki, R. (2024). Harnessing the power of digital resources in mathematics education: The potential of augmented reality and artificial intelligence. In S. Papadakis (Ed.), *IoT and ICT for educational applications*. Springer.

Marr, B. (2019, August 12). What is extended reality technology? A simple explanation for anyone. *Forbes*.

McMillan, K., Flood, K., & Glaeser, R. (2017). Virtual reality, augmented reality, mixed reality, and the marine conservation movement. *Aquatic Conservation: Marine and Freshwater Ecosystems, 27*, 162–168.

Microsoft. (2023, January 25). What is mixed reality?

Milgram, P., & Kishino, F. (1994). A taxonomy of mixed reality visual displays. *IEICE Transactions on Information and Systems, 77*(12), 1321–1329.

Moumoutzis, N., Paneva-Marinova, D., Xanthaki, C., Arapi, P., Pappas, N., & Christodoulakis, S. (2020). Using the PerFECt framework to establish an onlife community for theatre in mathematics to teach principles of computing. In W. K. Chan, B. Claycomb, H. Takakura, J.-J. Yang, Y. Teranishi, D. Towey, S. Segura, H. Shahriar, S. Reisman & S. I. Ahamed (Eds.), *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, 1084–1085, IEEE.

Oliver, M. (2011). Technological determinism in educational technology research: Some alternative ways of thinking about the relationship between learning and technology. *Journal of Computer Assisted Learning, 27*(5), 373–384.

Oliver, M. (2013). Learning technology: Theorising the tools we study. *British Journal of Educational Technology, 44*(1), 31–43.

Palmas, F., & Klinker, G. (2020). Defining extended reality training: A long-term definition for all industries. In M. Chang, D. G. Sampson, R. Huang, D. Hooshyar, N.-S. Chen, K. & M. Pedaste (Eds.), *IEEE 20th International Conference on Advanced Learning Technologies: Proceedings* (pp. 322–324). IEEE.

Pearl, H., Swanson, H., & Horn, M. (2019). Coordi: A virtual reality application for reasoning about mathematics in three dimensions. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–6.

Qushem, U. B., Christopoulos, A., Oyelere, S. S., Ogata, H., & Laakso, M. J. (2021). Multimodal technologies in precision education: Providing new opportunities or adding more challenges? *Education Sciences, 11*(7), 338.

Rokhsaritalemi, S., Sadeghi-Niaraki, A., & Choi, S.-M. (2020). A review on mixed reality: Current trends, challenges and prospects. *Applied Sciences, 10*(2), 636.

Scavarelli, A., Arya, A., & Teather, R. J. (2019, June 18–21). *Circles: Exploring multi-platform accessible, socially scalable VR in the classroom* [Paper presentation]. 2019 IEEE Games, Entertainment, Media Conference (GEM), New Haven, CT.

See, Z. S., Ledger, S., Goodman, L., Matthews, B., Jones, D., Fealy, S., Har Ooi, W., & Amin, M. (2023). Playable experiences through technologies: Opportunities and challenges for teaching simulation learning and extended reality solution creation. *The Journal of Information Technology Education: Innovations in Practice, 22*, 68–90.

Taber, K. S. (2020). Mediated learning leading development – the social development theory of Lev Vygotsky. In B. Akpan & T. J. Kennedy (Eds.), *Science education in theory and practice: An introductory guide to learning theory* (pp. 277–291). Springer.

Taber, K. S., & Li, X. (2021). The vicarious and the virtual: A Vygotskian perspective on digital learning resources as tools for scaffolding conceptual development. In A. M. Columbus (Ed.), *Advances in psychology research* (Vol. 143). Nova Science Publishers.

Thorpe, M. (2002). From independent learning to collaborative learning: new communities of practice in open, distance and distributed learning. In *Distributed Learning* (pp. 131–151). Routledge.

Tomaschko, M., & Hohenwarter, M. (2019). Augmented reality in mathematics education: The case of GeoGebra AR. In *Augmented reality in educational settings* (pp. 325–346). Brill.

Verenikina, I. (2010). Vygotsky in twenty-first-century research. In J. Herrington & B. Hunter (Eds.), *Proceedings of EdMedia 2010: World Conference on Educational Multimedia, Hypermedia & Telecommunication*. AACE.

Videla, R., Aguayo, C., & Veloz, T. (2021, September). From STEM to STEAM: An enactive and ecological continuum. In *Frontiers in Education, 6*, 709560. Frontiers Media SA.

Walkington, C., Nathan, M. J., Huang, W., Hunnicutt, J., & Washington, J. (2023). Multimodal analysis of interaction data from embodied education technologies. *Educational technology research and development*, 1–20.

Does ChatGPT make the grade?

Jude Brady (International Education Research Hub), **Martina Kuvalja** (Digital Assessment and Evaluation), **Alison Rodrigues** (International Education Research Hub) **and Sarah Hughes** (Digital Assessment and Evaluation)

This study took place in March 2023, just four months after OpenAI launched ChatGPT as a free service into the public domain. The generative artificial intelligence (AI) platform attracted over 100 million users in two months (Milmo, 2023). Its rapid growth and potential uses sparked a great deal of discussion about the implications for teaching, learning and assessment.

Our research is an early attempt from within Cambridge University Press & Assessment to explore the ways in which ChatGPT might be used in an assessment context. For this research, we engaged three undergraduate students to complete coursework essays with the assistance of ChatGPT. We selected a coursework task from the Cambridge IGCSE™ Global Perspectives syllabus because the syllabus encourages learners to think about and explore solutions to significant global issues. Students need to consider different perspectives and contexts. They also develop transferable skills to complement learning in other curricular areas.

We chose Component 2, the Individual Report (IR), which is a coursework component requiring candidates to write an essay. The skills assessed are:

- researching, analysing and evaluating information
- developing and justifying a line of reasoning
- reflecting on processes and outcomes
- communicating information and reasoning.

Component 2 requires that learners respond to tasks relating to the following topical issues of global importance:

- Belief systems
- Biodiversity and ecosystem loss
- Changing communities
- Digital world
- Family
- Humans and other species
- Sustainable living
- Trade and aid.

Given the broad range of skills required and breadth of topics available for the Global Perspectives Component 2, we considered that it lent itself well to our research task which required students to write essays with the assistance of ChatGPT. Through a comparison of the students' final essays with their ChatGPT outputs, analysis of ChatGPT chat logs, and interviews with students, the research explores the different extents to which students might use generative AI to help with essay-based coursework assessments. Findings from the work map the process that students adopted to navigate the ChatGPT platform and its outputs to arrive at a complete essay.

Findings from our work cannot be generalised to all possible uses of ChatGPT in essay writing due to the specific context and small sample size, but the broad remit of the work allows us to draw useful initial conclusions and identify areas for further research. The study highlights areas for consideration from a compliance and policy perspective; it invites discussion around the strengths and weaknesses of ChatGPT as an assessment aid and the limits of acceptable use. Additionally, when considering the future of assessment, questions can be raised about what skills are being measured and assessed when students engage with this type of technology in comparison to traditional methods. We hope that this early work will provide some ideas to help construct a holistic portfolio of research which examines and further explores the strengths and weaknesses of generative AI in essay writing and assessment.

Literature review

What is ChatGPT?

ChatGPT is a chatbot driven by a generative AI program. ChatGPT, like other generative AI chatbots such as Bing Chat, Microsoft Copilot and Bard, can generate novel outputs in response to prompts from the user, just like having a conversation. In practice, this means that a user can type in a question or instruction and the chatbot will generate a new response every time. Its responses are based on training data comprised of millions of websites, media outputs, journals, and books. The outputs are human-like and content is generated quickly. The chatbot is easily accessible provided the user has a device with an internet connection.

How does it work?

The ChatGPT program is based on a Large Language Model (LLM). Yosifova (2023) indicates that the model is "large" because it is informed by masses of data, and the model itself has many dimensions, layers, and connections or pathways between its different parts. In its training phase, the GPT-3.5 model developed and improved its ability to predict based on 175 billion parameters (*ibid*). The training sources were dated up to September 2021. Sometimes LLMs can be trained for a specific purpose, such as translation; however, ChatGPT uses a general model that aims to produce human-like language. As a result of the general training, the LLM is very powerful because it can be used for a huge range of tasks from chat and summarising materials to solving mathematics problems, and rewriting code.

GPT-3.5 training involved receiving human feedback in the form of rankings and accuracy ratings of its outputs. These human inputs helped to further the LLM's improvement.

Organisational approaches to generative AI

Cambridge University Press & Assessment is exploring a variety of LLM research. Areas which are being investigated include:

- production of texts for students at specified CEFR levels[1]
- content creation capability, including multiple-choice questions and quizzes (Galaczi, 2023)
- learning and assessment-focused applications and automarking capability.

Our work sits alongside these other investigations to offer some insight into the use of generative AI by students in an assessment context. The following review, drawn from academic literature and grey literature including blogs and opinion pieces, explores the perceived risks and opportunities of generative AI in this area.

Risks for assessment

Academic integrity

The most pertinent risk of generative AI to assessment is the potential for misuse. There are concerns that students could pass off AI-generated work as their own (Eke, 2023). Currently, AI detection tools are not reliable enough to be used to determine whether responses or answers to an assessment are partially or completely AI-generated, which means that students could be falsely accused of academic malpractice or indeed get away with cheating (Dalalah & Dalalah, 2023).

Reliability of information

There is evidence of ChatGPT generating false information and "hallucinations" which take the form of very plausible sounding references that do not exist (Dale, 2021; Perkins, 2023). If students are not trained to verify their sources, they will encounter challenges in distinguishing between facts and fabrications and possibly develop a knowledge base founded on fiction. Further to this, research has shown that GPT-3.5 is susceptible to different types of faulty reasoning (Marcus & Davis, 2020).

Inbuilt bias

Dwivedi et al. (2023) infer a risk when they describe how the information generated by AI could exhibit bias and privilege. The GPT model is largely trained on English language materials meaning these sources are not reflective of the diversity of views, perspectives and cultural truths that are prolific across the world (Lebovitz et al., 2023). Politically this is important because if AI is used to provide students with information and answers, to suggest ways of phrasing, mark assessments, author assessments, screen university or teaching applicants, or even to inform decision-making in education, it may well privilege and perpetuate one kind of perspective (e.g., a global north and white-centric viewpoint).

[1] The Common European Framework of Reference for Languages.

Legal concerns

There are unresolved complications around the use of generative AI in relation to copyright infringement and intellectual property rights (Dalalah & Dalalah, 2023; Lee, 2023). For example, if an assessment is authored mainly with the aid of generative AI, who owns the content? (Dippenaar, 2023). There are some instances of writers listing ChatGPT as a co-author on work (King & ChatGPT, 2023). However, high-profile journals such as *Nature* and *Science* have rejected this practice and will not accept chatbots as authors (Stokel-Walker, 2023).

Ethical concerns

Wider ethical concerns lie alongside these legal questions. Such concerns are not yet satisfactorily resolved because the extent to which generative AI has the capacity to cause harm is not fully understood and ethical frameworks for ChatGPT use are in development (CMS/W, 2023; Dwivedi et al., 2023).

Opportunities for learners, teachers, and assessors to exploit AI

Despite the risks, generative AI provides a set of unique assessment opportunities. Not only could such chatbots be used to help students learn, but they could also be used to author assessments and mark them.

Efficiency for teachers and students

Given the range of potential uses, the popularity of generative AI with educators is unsurprising. A survey of teachers in November 2023 suggests that 42 per cent of teachers are using AI to help with their work. Generative AI has the potential to improve efficiency for teachers if they use it with an awareness of its limitations. Kasneci et al. (2023) suggest that teachers could use chatbots to produce a text or model answer or to generate discussion prompts or lesson ideas, although these outputs would need reviewing by a human expert in the discipline. However, this could potentially save teachers time because the time spent authoring materials from scratch or searching online for appropriate teaching and learning supports could be reduced.

For students, ChatGPT and its equivalents could be used to provide a starting point for their research into a given and unfamiliar topic. The production of AI-generated content also opens the possibility of teachers and students reviewing these outputs together with a critical lens that invites discussion around the reliability, ethics, and efficacy of AI in education. It should be noted, however, that this kind of meta-reflection on the uses of generative AI introduces a new component into the teaching and learning arena. Educators would need to consider if or how courses and assessments should be adapted to accommodate and recognise learning which has taken place using generative AI.

Improved personalisation

Chatbots such as ChatGPT offer a unique opportunity for highly personalised learning. School students can learn through chatbot generated quizzes, summaries, and step-by-step explanations of how to solve specific problems (Kasneci et al., 2023). ChatGPT's "Socratic mode" allows students to be guided towards understanding through the Socratic questioning method.[2] Despite this potential, recent research into the Socratic mode's teaching of physics concepts found that the bot is unreliable at correcting misconceptions (Gregorcic & Pendrill, 2023). The authors suggest that the chatbot may prove more useful in producing erroneous explanations which students can then correct.

Availability and accessibility

One of the perceived advantages of generative AI is that there is a variety of models available (some free, some paid for), which can be accessed anytime and anywhere with an internet connection and suitable device. Students have the possibility of a learning dialogue with a chatbot without having to wait for peer or teacher assistance. For example, students could use ChatGPT as a revision aid and chatbots could assist learners who have additional needs (Kasneci et al., 2023). Finally, students who are learning in a language other than their first language may benefit from the instant translation, paraphrasing and clarification possibilities offered by ChatGPT.

Methodology

Research question

The research answers the following question:

- How do students use ChatGPT in essay writing?

The question was addressed from two angles. Firstly, we quantified the extent to which the students relied on ChatGPT-generated outputs to form the content of their essays. Secondly, we analysed interview data to gain insight into *how* students interacted with ChatGPT and their process of engagement with the technology for the set task. Students also shared their perceptions of the strengths, weaknesses, and purpose of ChatGPT in assessment.

Task

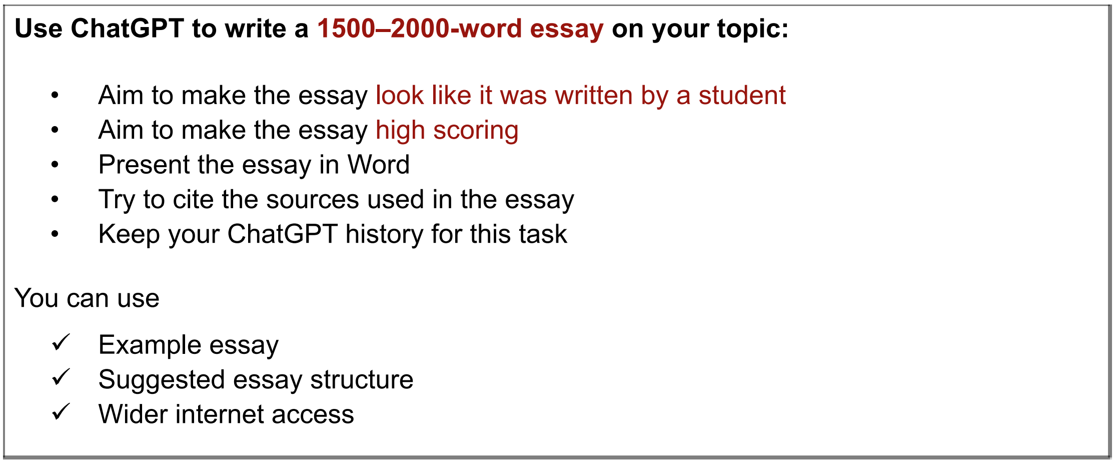
The focus of the study was to qualitatively explore how undergraduate students used ChatGPT technology to support them in a written assessment. Three undergraduate university students were invited to write two essays each for the Cambridge IGCSE Global Perspectives Individual Report (IR). Convenience sampling was used to select the students, and all students were reimbursed for expenses and paid for their time. An IGCSE Global Perspectives assessment task was selected because of the wide range of skills demanded. Furthermore, the assessment task requires students to gain (through research) a broad understanding of a topic of which they were expected to have limited prior knowledge. As the assessment is intended for IGCSE candidates, it was considered that the undergraduates would already have a good command of the skills required to engage effectively and meaningfully with the task, but that they would not have an in-depth knowledge of the topic areas. For these reasons, we expected that the undergraduates would be able to engage with the assessment task with relative ease and we could retain the research focus on their uses of ChatGPT to aid with essay writing.

[2] Socratic questioning is a shared dialogue by teachers, or in this case the chatbot, posing thought-provoking questions. The students then engage by asking their own questions. The discussion continues back and forth.

Although we chose to conduct the study with undergraduate-level students, the typical IGCSE student is 15–16 years old. It is likely that students of different age groups and/or with different levels of education would engage with both the technology and the process of essay writing in a different way to the undergraduates included in this research. Undergraduate students are also likely to be more skilled and experienced in essay composition and research-related tasks than their IGCSE counterparts. With these comments in mind, it should be noted that the findings from this qualitative and explorative study are not intended to be transferable or suggestive of the behaviours of wider populations. It intends to present a qualitative analysis of the practices of three undergraduate students who were provided with access to ChatGPT in an artificial assessment set-up.

The students had access to the ChatGPT "premium plan" to enable reliable access. They also had the choice of using either the version based on GPT-3.5 or the more recent GPT-4.0 version. Syllabus familiarisation training was provided. In this training, students gained an overview of the syllabus and its requirements, the Assessment Objectives and marking criteria, and they looked at an exemplar essay. After the familiarisation training, the students were invited to select two essay titles from 13 options that had been randomly selected in advance by the researcher. The titles were selected from genuine IR assessment titles submitted in November 2019.

Students were provided with the instructions shown in Figure 1.



Use ChatGPT to write a 1500–2000-word essay on your topic:

- Aim to make the essay look like it was written by a student
- Aim to make the essay high scoring
- Present the essay in Word
- Try to cite the sources used in the essay
- Keep your ChatGPT history for this task

You can use

✓ Example essay
✓ Suggested essay structure
✓ Wider internet access

Figure 1: Instructions provided to research participants

To avoid a scenario where the students deliberately authored poor essays, believing that these were reflective of the level of the typical (IGCSE) student, we asked them to try and make the essays "high scoring". As per the syllabus requirements, the essays needed to be presented in Microsoft Word and sources cited in a consistent way. As Figure 1 shows, the students were invited to use the training materials and internet as well as their prior knowledge. They were told that their approach was their decision and the researcher explained that the task was purposefully not over-prescriptive because we were interested in *how* they used the generative AI technology. It was also for this reason that the students were not provided with ChatGPT familiarisation training, although they each reported prior awareness of the technology.

Document comparison

Students retained the transcripts of their interactions with ChatGPT. These were submitted to the research team alongside the essays. An “overall plagiarism” percentage was calculated by comparing the ChatGPT transcripts to the students’ final essays using Copyleaks’ document comparison tool: “text compare”. The two text types (the essay and chatlog) were input into Copyleaks’ “text compare” for the tool to output a plagiarism percentage score. The score denotes how much of the essay had been copied and pasted or adapted from the ChatGPT outputs.

Copyleaks’ developers state that the “text compare” tool works by using “advanced algorithms” which “[look] for matches within the submitted text” (Jacob, n.d.). The explanation continues to outline how it uses:

> “lexical analysis, semantic analysis, and machine learning [...to...] uncover even subtle instances of plagiarism [...]. The algorithm then does a deep-dive, using fuzzy matching to uncover patterns and stylometry to check for differences in writing style.” (Jacob, n.d.)

The output documents highlight which sections of the student’s essay have been flagged as which type of plagiarism. The identification of paraphrasing in particular could be more sensitive than a human evaluator, although to the best of our knowledge there are no publicly available studies comparing the “text compare” identification of paraphrasing with that of humans.

The term “plagiarism” is used throughout this article, although the extent to which each of the students engaged in academic malpractice is debatable. Further to the overall plagiarism percentage score, the analyses output a percentage score for each of three levels of plagiarism: “identical” where sentences or phrases were lifted word for word from ChatGPT’s outputs; “minor amendments” where students made small changes to the ChatGPT content, and “paraphrased” content. The researcher compared the overall plagiarism score for each essay and student and made a qualitative and relative judgement between students as to whether each undergraduate had relied on ChatGPT to a “high”, “medium” or “low” extent when constructing the essays.

Interviews

Upon the submission of their essays, the students were interviewed by one of three researchers about their experience of using ChatGPT in writing their essays. The purpose of the interviews was to gain insights into the process of using ChatGPT in producing essays in contexts where students had little previous knowledge of the topic.

The one-to-one semi-structured interviews took between 40 minutes and 1 hour. Overall, the students were asked (i) how they had used ChatGPT to write essays, (ii) how they had integrated ChatGPT-generated content into their essays, and (iii) how well they thought ChatGPT had helped with their essay writing. The interview protocol was followed and some follow-up clarifying questions were asked by researchers, as appropriate. The interviews were audio recorded and transcribed using Microsoft Teams. The transcripts were then anonymised, edited, and analysed by applying thematic analysis. This approach included reviewing the transcripts, indexing segments into categories and, finally, identifying common themes across the interviews.

Ethical considerations

This research followed the British Educational Research Association's guidelines for conducting educational research. The students gave their written and verbal consent to participate in the research study and they were provided with opportunities to ask questions and to withdraw from the study. The identities of the students are obscured throughout the article by the use of pseudonyms and gender-neutral pronouns (e.g., "they" to refer to an individual).

Findings

To what extent did students rely on ChatGPT-generated outputs to form the content of their essays?

Following the Copyleaks analysis, it appeared that the students had taken three different approaches to the use of ChatGPT in their essay-writing tasks. When looking at the overall plagiarism percentage scores, Kim had the highest level of plagiarised content (with plagiarism scores of 70 per cent and 64 per cent for their essays). This finding indicates that Kim had interpreted the task in such a way that they relied heavily on the chatbot for essay content. Relative to Kim and Ronnie, Charly engaged in a "medium" level of overall plagiarism (44 per cent and 41 per cent), suggesting a relative mid-level of reliance on ChatGPT in constructing the essays. With Ronnie there was no evidence that they had copied and pasted (0 per cent) or minorly amended (0 per cent) text from the ChatGPT chatlog. Furthermore, there was very little evidence that ChatGPT generated content had been paraphrased (7 per cent and 4 per cent).

Table 1: Copyleaks' plagiarism scores for the ChatGPT-assisted essays

Essay	Student	Plagiarism overall	Identical	Minor changes	Paraphrase
Religion and conflict	Kim	70%	21%	17%	32%
Animal rights*	Kim	64%	13%	29%	22%
Legalising abortion*	Charly	44%	17%	12%	15%
Celebrities as role models	Charly	41%	2%	7%	31%
Capital punishment	Ronnie	7%	0%	0%	7%
Sweatshops	Ronnie	4%	0%	0%	4%

** Essay written with ChatGPT-4 instead of ChatGPT-3.5*

We interpret these different levels of overall plagiarism scores as indicating different levels of reliance on ChatGPT for essay writing in this task. In the findings below these approaches are referred to as "high", "medium", and "low" dependence, and this is a relative judgement made by the researchers by comparing the students' essays, plagiarism scores, and interviews.

Low dependence

Ronnie, who adopted the low plagiarism approach, was sceptical about the idea that ChatGPT alone, without human editing and input, could achieve good grades at university level. This scepticism is perhaps reflected in their approach to the task. As Table 1 illustrates, Ronnie did not simply plagiarise from ChatGPT. Instead, they developed the essay argument, and then used ChatGPT to elicit sources of information to vindicate their ideas:

> Ronnie: "[A] lot of the [essay] ideas just [depended] on the way that I wanted to steer it, and where I wanted to take it. Obviously, I was going to argue that [capital punishment] is not ethical because that's what I believe. So, I'd steer [ChatGPT] down that road …"

Here Ronnie explains that the essay argumentation and ideas were their own. Before engaging with the technology, they had decided that they would argue that capital punishment was unethical. Accordingly, they "steered" ChatGPT to provide information that was useful to the construction of this stance. To elicit the desired information, Ronnie needed to alter their prompts. They explained that ChatGPT's inbuilt safeguards initially mean that the system refused to answer their questions:

> Ronnie: "First of all it [ChatGPT] didn't like really want to register the word 'death' […] I switched to 'capital punishment' and then it went and started giving me a more detailed answer …"

Ronnie rapidly found a way around ChatGPT's safeguarding mechanisms by avoiding the word "death", and this allowed them to gain information relevant to the case that they wanted to argue. Although Ronnie later claimed that ChatGPT "doesn't really have an opinion", they found ways to force it to mimic an opinion, for example by asking the system to give a perspective from a particular standpoint (such as "from the perspective of someone that lived in Bangladesh" for the essay question about the ethics of "sweatshops").

Medium dependence

Charly adopted a middle approach with 41–44 per cent of their essays plagiarised in some way. Where they had adopted a "copy and paste" approach, they were concerned that this was at odds with the rest of their essay's style:

> Charly: "There were very rare moments where I just copied and pasted straight from GPT […] at the points where I did, it didn't feel real. It felt like a fact file rather than an essay."

Charly may have objected to the "fact file" style for aesthetic reasons and considered that the ChatGPT voice was ill-suited to an essay of this type. Whatever Charly's reasons, they indicated a preference for expressing content in their own words. This preference is partly borne out in the findings displayed in Table 1. For the essay about celebrities, Charly's use of "copy and paste" was limited to just 2 per cent, but they paraphrased a further 31 per cent of their essay material from ChatGPT's outputs. In the essay about legalising abortion, however, 17 per cent of the overall essay was copied directly from the ChatGPT chatlog and 15 per cent was paraphrased. It is unclear why Charly adopted these slightly different approaches to ChatGPT use across the two essays. It could be that by the time Charly started work on their second essay (the legalising abortion essay), they were running short on time and so resorted to a higher degree of direct dependence on ChatGPT's outputs.

High dependence

Kim's essays were constructed more heavily from ChatGPT outputs compared to both Ronnie and Charly. Despite plagiarising between 64–70 per cent of the ChatGPT log, Kim noted that a complete "copy and paste" approach would be unlikely to be successful:

> Kim: "Overall ... [ChatGPT is] not great [for essay writing]. [...] you can't just [...] ask it to write it and then copy and paste it. It just wouldn't work. So I think, it definitely takes more work than you think it would initially do. I didn't think it was going to take 4 hours to write an essay with GPT, whereas it does."

Kim was unimpressed with the idea of writing an essay with only ChatGPT. In line with the other two students, they felt that this "wouldn't work". Even though Kim's essays included the greatest proportion of content plagiarised from the ChatGPT transcript, they undertook "more work" than they initially had anticipated would be necessary. Much of this work was to do with the selection and synthesis of information, which was also remarked on by Charly.

Kim's interview indicated that they found the process of using ChatGPT unfulfilling. When asked about their perception of the quality of their essays, Kim responded:

> Kim: "I literally have done no work for this assignment. But somehow, I managed to do OK. [...] it was like copying and pasting and just more like trying to find the sources where they got the information from rather than just like you know, researching it yourself. I don't know — I think it [that] it didn't feel as momentous, you know?"

They felt dissatisfied with the process of essay writing with ChatGPT because they perceived that they had "done no work for this assignment", and they viewed their input largely in terms of locating sources and "copy and pasting". Kim may have perceived the skills required to elicit information from ChatGPT, verify sources, and select and synthesise ChatGPT-generated content as less valuable than those needed to "[research] it yourself". As a result, Kim felt underwhelmed rather than pleased with the "momentous" achievement of having researched, processed, synthesised, structured, and authored high-quality work. Kim's comments open a further question about the way in which engagement with ChatGPT affects the constructs measured in a coursework assessment and the motivation and satisfaction of learners.

How did students interact with ChatGPT in the process of essay writing?

Each of the students used ChatGPT-generated content in different ways and to different extents. Nonetheless, the analysis found similarities in their process of writing essays while interacting with this technology. This process (shown in Figure 2) comprised of:

1. orientation
2. specific enquiries
3. occasional verification
4. structuring; and
5. writing up.

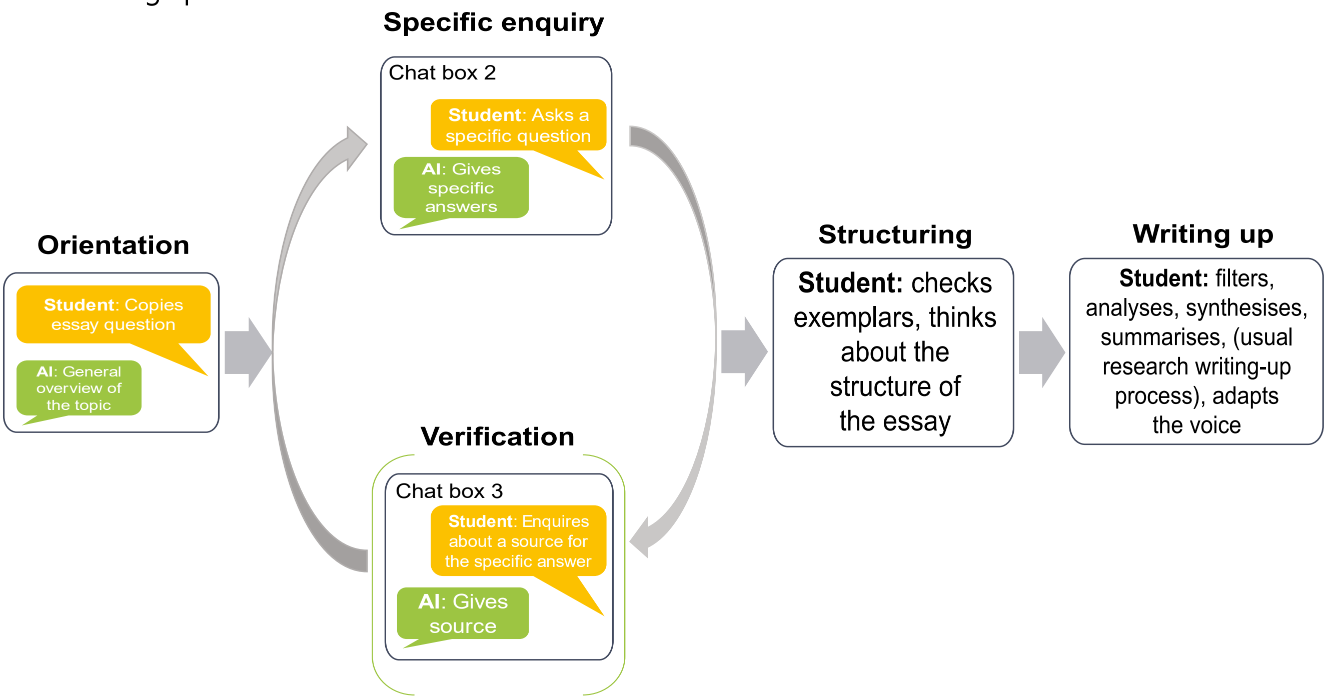

Specific enquiry
Chat box 2
Student: Asks a specific question
AI: Gives specific answers
Orientation
Student: Copies essay question
AI: General overview of the topic
Structuring
Student: checks exemplars, thinks about the structure of the essay
Writing up
Student: filters, analyses, synthesises, summarises, (usual research writing-up process), adapts the voice
Verification
Chat box 3
Student: Enquires about a source for the specific answer
AI: Gives source

Figure 2: Students' essay-writing process using ChatGPT. The verification step is presented in brackets to show that students did not consistently verify information.

Orientation and specific enquiries

Students reported starting off by entering the essay question into the ChatGPT chat box or asking a general question about the topic ("orientation"). It helped students, who did not know much about a topic, to get a general overview of it. This was followed by multiple "specific enquiries"; students would start multiple chats or enter new questions into the same chat to enquire about specific aspects of the topic (Figure 1):

> Charly: "So, I'd start new chats [...] the pattern I followed was: I'd ask it a general question like [what is the] Christian debate on abortion, for example, and it would give maybe four or five points. And then after that, I'd ask more with details, facts and statistics."

Students were, overall, quite impressed with the speed of access to information through ChatGPT. They felt that the purpose of ChatGPT was very similar to the purpose of any other internet search engine – to provide information, just much quicker and more user-friendly:

> Ronnie: "It's just [...] a better Wikipedia. It's like times... infinity better [than] Wikipedia."

Verification

Occasionally, students attempted to verify the ChatGPT-generated content ("verification") by either asking ChatGPT to provide the source or by searching for the sources without the help of ChatGPT by using internet search engines. The students had a strong sense of ethics around verification of the AI-generated content and referencing, but when asked about verification practices and referencing used in their essay writing, they reported they often skipped this part of the process. This was due to the difficulty of verifying the content, time constraints and a lack of understanding around the expected standards of verification and referencing:

> Kim: "[...] we didn't have time to [...] verify it. I probably should do that because it's probably [...] I don't know how reliable it is [...] it doesn't specifically give you sources when it gives you the information. You can't do that for every piece of information that you're gonna get, because otherwise, it's just gonna go on."

Tracking down the source of ChatGPT's information was frustrating. Kim felt that it "would have been better" if they had found and referenced the sources in the first place, rather than relying on ChatGPT and later trying to verify its information, which is now possible (e.g., see Microsoft Copilot).

Ronnie, on the other hand, used ChatGPT output and "searched on Google Scholar" to find "text that link[ed]" to the ChatGPT output. They would then reference the Google Scholar source. With this method of searching for and then citing likely-sounding sources in the essay, Ronnie found that all the output they wished to include was "quite supportable" mainly because there was "nothing outrageous" to justify.

Structuring

At the next "structuring" stage, the students gathered all the content that ChatGPT generated and started thinking about the form of their essays. They reported checking the exemplar essay to gain an idea of an appropriate format and content type. Ronnie and Kim started to structure their essays after the orientation stage, whereas Charly first gathered 22 pages of information from ChatGPT before starting to structure the essay. The structuring process was not contained or linear for any of the students and all showed evidence of returning to the stage at different points in the essay-writing process. In the structuring stage, the students reported difficulties in obtaining appropriate introductions and conclusions from ChatGPT:

> Charly: "I messed around with trying introductions and conclusions. And they were very poor because [ChatGPT] doesn't come to a conclusion."

Writing up

In this stage, students needed to decide how to write up the essay so that it had a coherent argument and followed the component's suggested essay structure.

This process was complex and required higher-level critical thinking skills, including analysis, synthesis, and evaluation of the ChatGPT-generated content (and students were aware of that):

> Charly: "[Using ChatGPT is] not that bad because you can then apply your research skills and select and synthesise. You will get a very low mark just [by] using ChatGPT. [...] You need to have [...] the skills developed [...] but then I feel like those skills are developed from not using a source like ChatGPT [...] it's [...] a paradox."

Integrating ChatGPT-generated content into an essay required an adaptation of style and voice. The ChatGPT outputs were often deemed unsuitable for copying verbatim into the essays because they lacked "style" and were "too logical":

> Kim: "You make it more so that it sounds like it's more appropriate for that essay, as your own voice [...] because [...] it was very directive... it's more third person [...] It's more of a telling rather than like you're explaining."

As a further observation, the students did not use the entire set of functionalities offered by ChatGPT (e.g., restructuring, copyediting, proofreading, and providing feedback). This could be due to the lack of training and time pressures. Finally, at the time of this study, ChatGPT had been widely available for only four months, which means that students may not have had sufficient time to familiarise themselves with all its features. It is reasonable to presume that students' general familiarity with ChatGPT will be significantly more advanced by the time this journal article is published.

Conclusion

As suggested throughout the article, there are several limitations that must be recognised in the interpretation of the data and presentation of findings. Unlike a naturalistic setting, the research participants were explicitly asked to use ChatGPT to write their essays. As such, they may have engaged with generative AI to a greater extent than if they had been genuinely studying towards this qualification. Secondly, the research participants were undergraduates aged 18–22, and it should not be assumed that students in this category reflect the behaviours of a younger cohort who may have a different approach and skill level when it comes to engaging with generative AI. Thirdly, the undergraduate students had little previous knowledge of the essay topics, and had only two days to produce the essays. These factors may have affected their approach to the process of essay writing compared to a naturalistic setting.

Despite the limitations, findings from this research provide an indication of how the selected students engaged with ChatGPT and offer insight into their perceptions of the utility and ethics of using such a tool to assist with essay writing in an assessment context. Notably, despite different levels of reliance on ChatGPT, the students used the technology in a similar way: primarily as an information gathering and producing tool. There was limited evidence of them exploiting ChatGPT to its full potential, as an editor, proofreader, or to provide formative feedback – for example. As previously noted, this seemingly limited awareness of ChatGPT's potential may be because the students had not explored it in depth prior to the research task and did not have time to test it or be creative with it during the essay-writing task.

The students understood that ChatGPT generated both accurate and false (and outdated) information, and they did not always verify the information provided to them. They recognised this as a problem and suggested that ChatGPT's overall capabilities and outputs were not of a high enough standard to facilitate top marks in an essay at IGCSE level or above. As such, they may not have deployed ChatGPT's full suite of uses because they did not believe that these would add value to their essays. Had the students been provided with ChatGPT familiarisation training or with the addition of exploration time prior to starting the essay-writing tasks, they might have uncovered the technology's capabilities, used it in more varied ways or been more impressed by its functions and applications. Similarly, had the LLM output included the sources behind its content (for example, see Microsoft Copilot), the students might have interacted with it in a different manner and had more confidence in using the AI-generated content.

As the technology evolves and as users become more accustomed to its potential applications and uses, students such as those in our study sample may develop into more skilled users of generative AI and they may perceive that it can, in fact, outperform humans in tasks such as essay writing at IGCSE level.

As well as the students' perceptions, this research has highlighted the importance of higher-order thinking skills for AI-assisted essay writing, and that is unlikely to change any time soon. The students reported challenges in using ChatGPT for content generation because they could not easily verify the chatbot's outputs, nor did they find ChatGPT's default voice to be appropriate to their task. These findings accord with current wider claims that ChatGPT does not necessarily excel in these areas (University of Cambridge, 2023). Given that students with lower academic performance often also display poor critical thinking (Behrens, 1996; Fong et al., 2017) and poor metacognitive skills (Pintrich & De Groot, 1990; Young & Fry, 2008), it would be fair to assume that low-performing students in particular could be potentially negatively affected by using ChatGPT for content generation.

With a view to the future, other research in the area of AI and assessments has mapped out what is possible and what is desirable (Abu Sitta et al., 2023). Such research explores how to capitalise on AI in such a way that it enhances rather than diminishes human capabilities. Future-oriented research combined with research about current practice and engagement in AI can help to inform institutional policies and guidelines which are under continual review, given the fast-developing nature of the area. It could be useful to include undergraduate students' voices in the design of such policies and guidelines because, as the students in this study have shown, they may have valuable perspectives on what ethical and legitimate use of generative AI could look like in an assessment context.

References

Abu Sitta, F., Maddox, B., Casebourne, I., Hughes, S., Kuvalja, M., Hannam, J., & Oates, T. (2023). *The futures of assessment: Navigating uncertainties through the lenses of anticipatory thinking*. DEFI & Cambridge University Press & Assessment.

Behrens, P. J. (1996). The Watson-Glaser critical thinking appraisal and academic performance of diploma school students. *Journal of Nursing Education, 35*(1), 34–36.

CMS/W. (2023, January 13). *Advice and responses from faculty on ChatGPT and A.I.-assisted writing*. MIT Comparative Media Studies/Writing.

Dalalah, D., & Dalalah, O. M. A. (2023). The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT. *The International Journal of Management Education, 21*(2), 100822.

Dale, R. (2021). GPT-3: What's it good for? *Natural Language Engineering, 27*(1), 113-118.

Dippenaar, B. (2023, June 12). *ChatGPT and AI: Navigating uncharted copyright territory*. Lexology.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71*, 102642.

Eke, D. O. (2023). ChatGPT and the rise of generative AI: Threat to academic integrity? *Journal of Responsible Technology, 13*, 100060.

Fong, C. J., Kim, Y., Davis, C. W., Hoang, T., & Kim, Y. W. (2017). A meta-analysis on critical thinking and community college student achievement. *Thinking Skills and Creativity, 26*, 71–83.

Galaczi, E. (2023). *English language education in the era of generative AI: Our perspective*. Cambridge University Press & Assessment.

Gregorcic, B., & Pendrill, A.-M. (2023). ChatGPT and the frustrated Socrates. *Physics Education, 58*(3), 035021.

Jacob, S. (n.d.). Copyleaks Plagiarism Review – Originality.AI. Copyleaks. Retrieved 8 January 2024, from https://originality.ai

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274.

King, M. R. & ChatGPT. (2023). A conversation on artificial intelligence, chatbots, and plagiarism in higher education. *Cellular and Molecular Bioengineering, 16*(1), 1–2.

Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2023). The No. 1 question to ask when evaluating AI tools. *MIT Sloan Management Review, 64*(3).

Lee, J. Y. (2023). Can an artificial intelligence chatbot be the author of a scholarly article? *Journal of Educational Evaluation for Health Professions, 20*(6).

Marcus, G., & Davis, E. (2020, August 22). *GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about.* MIT Technology Review.

Milmo, D. (2023, February 2). ChatGPT reaches 100 million users two months after launch. *The Guardian*.

Perkins, M. (2023). Academic integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice, 20*(2).

Pintrich, P. R., & De Groot, V. E. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*, 33–40.

Stokel-Walker, C. (2023). ChatGPT listed as author on research papers: Many scientists disapprove. *Nature, 613*(7945), 620–621.

University of Cambridge. (2023). *ChatGPT (We need to talk)*.

Yosifova, A. (2023, June 28). *ChatGPT: How to understand and compete with the AI bot*. 365 Data Science.

Young, A., & Fry, J. D. (2008). Metacognitive awareness and academic achievement in college students. *Journal of the Scholarship of Teaching and Learning, 8*(2), 1–10.

How do approaches to curriculum mapping affect comparability claims? An analysis of mathematics curriculum content across two educational jurisdictions

Nicky Rushton (Cambridge University Press & Assessment) **Dominika Majewska** (Cambridge University Press & Assessment) **and Stuart Shaw** (Institute of Education, University College London)

Curriculum mapping is a method used within comparability studies to make comparisons of curriculum content within multiple settings: usually multiple jurisdictions or multiple specifications. These maps form the first part of the comparability studies. They present information from the jurisdictions/syllabuses (such as features of the education system or areas of curriculum content) in tables to make it easy for experts to make comparisons across the jurisdictions/syllabuses. These comparisons provide the evidence for claims about the jurisdictions/specifications. For example, the Department for Education (2012) used its mapping of curriculum content from six jurisdictions to claim that "Some mathematics curricula of high-performing jurisdictions are much more challenging than the 1999 and 2007 national curriculum for England, in particular on *number* and *algebra*, though *data and statistics* is slightly more challenging in England" (p. 3).

Curriculum maps are often used to compare the breadth and depth of curricula or specifications for qualifications (e.g., Alcántara, 2016; Department for Education, 2012; Ofqual, 2012). They often include the aims and content of the curriculum/specification, and features of examinations based on the curricula/specifications. Additional maps are sometimes included to provide information about the context, which enhances the analysis and key features of the education systems. Maps have also been used to compare features of interest across different jurisdictions (e.g., Elliott, 2014). Although curriculum maps have been published in policy documents and reports, studies using this method are rarely published and very little has been written about it in the academic literature (Elliott, 2014; Greatorex et al., 2019).

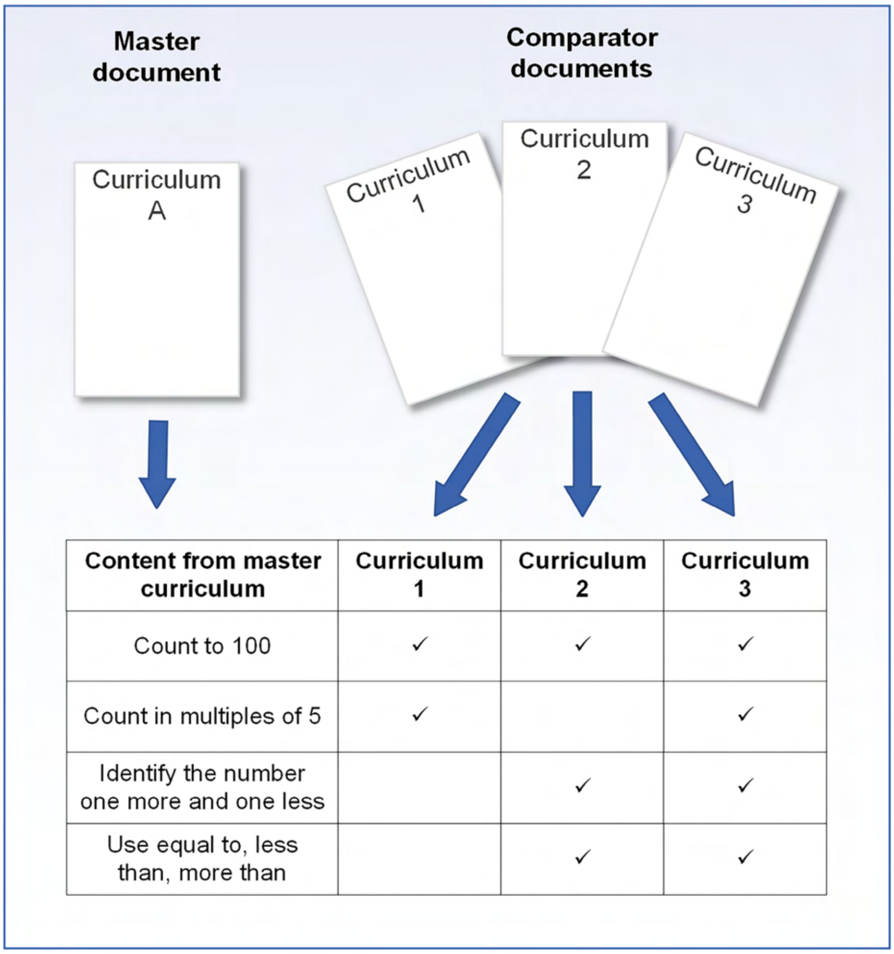
At this point, it may be helpful to clarify what is meant by the term *curriculum* in the context of curriculum mapping. We use the term curriculum to describe any document which forms part of the intended curriculum[1] in its respective jurisdiction. These can include:

- syllabuses or specifications, which set out the structure and content of courses and assessments
- educational standards, which are the documents used in the US to describe what students should "know and be able to do" (paragraph 2, Common Core State Standards Initiative, 2022).

It is important to note that comparisons that are based on documents which define the intended curriculum cannot provide any information about other types of curricula, such as the taught curriculum or the learned curriculum. Nor can they provide any information about the way in which the subject is taught within classrooms.

The maps usually consist of comparison tables or spreadsheets with specific comparators within each column (e.g., qualifications) and particular information in the rows (e.g., curriculum content) (Elliott, 2014). They differ from simply recording information as they enable direct comparisons to be made between jurisdictions by reading across a row; therefore, they are a tool to inform thinking and enable judgements. A document known as the *master* curriculum (Elliott, 2014) is always used as the basis of the comparison. Content from the other curricula (the *comparators*) is matched to this master curriculum, as can be seen in the examples of content mapping shown in Figure 1. Curriculum maps use one or more symbols in each cell of the table to indicate whether content from the master curriculum is covered in the comparators. For example, the TIMSS topic trace mapping (see Schmidt et al., 2018) uses two symbols to show whether each topic is taught in a particular year group and whether there is a particular focus on that content area in that year. Alternatively, maps may contain content descriptions instead of symbols, so that they can provide more detailed information.

[1] The intended curriculum is "the overt curriculum that is acknowledged in policy statements as that which schools or other educational institutions or arrangements set out to accomplish" Kridel, C. A. (2010). Intended curriculum. In *Encyclopedia of curriculum studies* (Vol. 1, pp. 179-181). Sage Publications.



Content from master curriculum	Curriculum 1	Curriculum 2	Curriculum 3
Count to 100	✓	✓	✓
Count in multiples of 5	✓		✓
Identify the number one more and one less		✓	✓
Use equal to, less than, more than		✓	✓

Figure 1: Example of a curriculum map

There are circumstances when the mapping process needs to be altered slightly. For example, it may not be possible to map all the content because of time or budget limitations. There do not appear to be any studies in the public domain which consider a sub-set of content, so it is not possible to ascertain how this reduced content would affect any conclusions that could be drawn from the mappings.

More commonly, the structure of the content across the curricula may affect the mapping. For example, the content may be arranged differently across age groups in the curricula being compared. An example of this is the Department for Education (2012) mapping, where some curricula set out content by single year groups (e.g., Singapore) while others had multiple year group spans (e.g., Massachusetts and Finland). The authors found that this difference made it technically challenging to carry out the mapping and difficult to identify differences in the sequencing of content.

In this article, we will use a mapping study comparing the Common Core State Standards (CCSS) in the United States (US) and mathematics national curricula in England to discuss approaches to mapping when a sub-set of content is used or when curricula are structured differently by age. We will also discuss how the approaches differ from mapping whole curricula with matching age structures in terms of the conclusions or comparability claims that can be drawn from them.

Overview of the curricula

We used the following documents for our comparison:

- the CCSS for mathematics for grades K to 8, and
- the mathematics programmes of study for key stages (KS) 1, 2 and 3 (years 1–9).

These ranges of grades/years are considered equivalent (the grade number in the US is one less than the equivalent school year in England). We chose this range of grades because the CCSS standards are only aligned to particular grades until the end of grade 8. Beyond that, the standards are allocated to content areas, making it impossible to compare when content was taught. Additionally, year 10 in England marks the point when the curriculum differentiates between the content that is taught to all students and the content that is only taught to higher attaining students. This would complicate comparisons with the CCSS as it would require separate analysis of the content for all students and the content for higher achieving students.

The CCSS (see NGA Center & CCSSO, 2010)

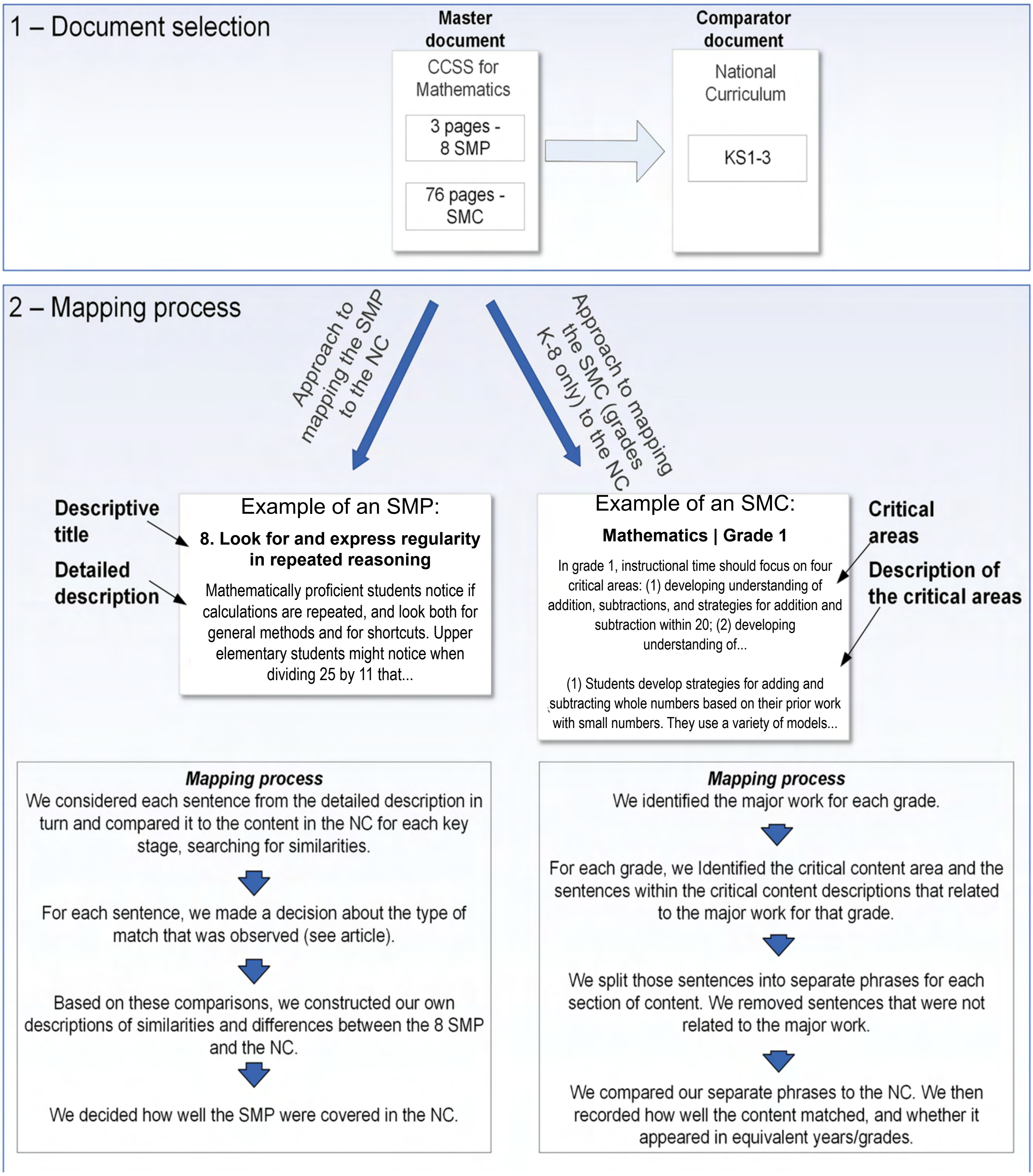
The CCSS in the US are "a set of high-quality academic standards in mathematics and English language arts/literacy (ELA). These learning goals outline what a student should know and be able to do at the end of each grade" (Common Core State Standards Initiative, 2022, para 2). Use of the CCSS is not compulsory, but many states have chosen to adopt it, or have based their own standards on it. The CCSS for mathematics document is divided into two parts: the eight Standards for Mathematical Practice (SMP) and the Standards for Mathematical Content (SMC). The SMP are common to all grades and describe the expertise that teachers should aim to develop in learners (NGA Center & CCSSO, 2010). The SMC set out what students are expected to understand and do, and are set out by grade from kindergarten (K) to grade 8.

The national curriculum (see Department for Education, 2013a; Department for Education, 2013b)

The national curriculum in England "is a set of subjects and standards used by primary and secondary schools, so children learn the same things. It covers what subjects are taught and the standards children should reach in each subject" (UK Government, n.d., para 1) and is compulsory for many state schools in England. The documents contain a programme of study that lists the content that students should cover in particular key stages of schooling, and the matters, skills and processes that students are expected to be able to know and understand in those content areas (Department for Education, 2013b). These are set out by year group in KS1 and 2 (years 1–6), but KS3 content is common to all year groups (years 7–9).

Curriculum mapping methods

We used the CCSS as the master curriculum (Figure 1), because we wanted to see how its content differed from the national curriculum rather than the other way around. Carrying out a full-scale curriculum mapping comparison demands inordinate time and effort given the ultimate aims; therefore, we decided we could not map all the standards from the CCSS. Only three pages within the CCSS were devoted to the SMP, so we decided it would be possible to map that content. However, 76 pages were devoted to the SMC, so it was only possible to map a subset of the SMC content. The need to adopt different approaches for the two sections of the CCSS provided us with the opportunity to compare the two curriculum mapping methods (see Figure 2 for a visual representation of this and the mapping process).



Example of an SMP:

8. Look for and express regularity in repeated reasoning

Mathematically proficient students notice if calculations are repeated, and look both for general methods and for shortcuts. Upper elementary students might notice when dividing 25 by 11 that...

Example of an SMC:

Mathematics | Grade 1

In grade 1, instructional time should focus on four critical areas: (1) developing understanding of addition, subtractions, and strategies for addition and subtraction within 20; (2) developing understanding of...

(1) Students develop strategies for adding and subtracting whole numbers based on their prior work with small numbers. They use a variety of models...

Mapping process

We considered each sentence from the detailed description in turn and compared it to the content in the NC for each key stage, searching for similarities.

For each sentence, we made a decision about the type of match that was observed (see article).

Based on these comparisons, we constructed our own descriptions of similarities and differences between the 8 SMP and the NC.

We decided how well the SMP were covered in the NC.

Mapping process

We identified the major work for each grade.

For each grade, we Identified the critical content area and the sentences within the critical content descriptions that related to the major work for that grade.

We split those sentences into separate phrases for each section of content. We removed sentences that were not related to the major work.

We compared our separate phrases to the NC. We then recorded how well the content matched, and whether it appeared in equivalent years/grades.

Figure 2: Approaches to mapping the Standards for Mathematical Practice (SMP) and the Standards for Mathematical Content (SMC)

Approach to mapping the SMP

There is no overarching content for the whole national curriculum, which means that there is no direct equivalent of the SMP. However, the skills described in the SMP can be found throughout the national curriculum content for specific year groups in KS1 and 2 (ages 6–11) and in the working mathematically content in the KS3 national curriculum (ages 12–14).

For the curriculum mapping, we took each sentence within the detailed SMP descriptions and compared it to the content in the national curriculum for each year to identify similar content. We then decided whether there was:

- a complete match with identical content (✓)
- a partial match with some matching content found (~)
- no match (×).

Where we could not find a match, but we felt that the content was needed in order to teach content that was listed, we noted this in the "notes on implicit matches" column. Table 1 shows an extract from this mapping. Row 2 shows the descriptive title for the first SMP, and rows 3-5 show the first three sentences from the detailed description for that SMP. The KS1, KS2 and KS3 columns show the matches for each sentence of the SMP. The best match for each sentence of the SMP was recorded in the overall KS1–3 column (e.g., the best match for row 4 was the partial match found in KS2, so this was the level of match recorded in the overall column). Finally, we recorded overall level of matching for the descriptive title of each SMP by tallying the number of sentences that were coded with each type of match (see Table 1, row 2).

Using these comparisons, we were able to make judgements about how well each of the SMP was covered explicitly and implicitly in the national curriculum.

Table 1: Example of curriculum mapping between the SMP and the mathematics national curriculum for KS1–3

SMP	SMP detailed description	KS1	KS2	KS3	Overall KS1–3	Notes on implicit matches
1. Make sense of problems and persevere in solving them.	**N/A**	✓✓ ~ ××××××	✓✓ ~~~ ××××	✓✓✓ ××××××	✓✓✓✓✓ ~~ ××	
	Mathematically proficient students start by explaining to themselves the meaning of a problem and looking for entry points to its solution.	×	×	×	×	Students will have to do this but is not stated in the documentation.
	They analyze givens, constraints, relationships, and goals.	×	~	×	~	Not explicitly covered but is needed when solving problems.
	They make conjectures about the form and meaning of the solution and plan a solution pathway rather than simply jumping into a solution attempt.	×	~	✓	✓	-

Approach to mapping the SMC

Because we were not able to map the whole of the SMC for grades K–8 as we normally would in comparisons, we had to reduce the content that was included. Therefore, we decided to focus on content associated with the major works –

the most important content for each grade which was intended to receive the majority of the teaching time (Achieve the Core, n.d.). There are five major works:

- "Addition and subtraction" (grades K–2)
- "Multiplication and division of whole numbers and fractions" (grades 3–5)
- "Ratios and proportional relationships, and early algebraic expressions and equations" (grade 6)
- "Ratios and proportional relationships, and arithmetic of rational numbers" (grade 7)
- "Linear algebra and linear functions" (grade 8)
(NGA Center & CCSSO, n.d., section 1).

While previous studies do not appear to have used a sub-set of content in this way, we thought the method would provide us with useful information about the differences between the mathematics curricula in the two countries.

For the SMC mapping we compared the phrases we had identified in the descriptions to the national curriculum to identify matching content. We used ticks and tildes to show which year groups in England contained matching or partially matching content, and light grey shading to indicate the equivalent grades/school years. Table 2 rows 3–6 show the result of these comparisons for the K–2 major work, "Addition and subtraction". We then summarised the mappings for the major work row (shown in bold in row 2). Note that the grade 2 SMC descriptions are not shown in Table 2 but our curriculum mapping showed that it was also taught in year 3, hence the tick in row 2 for year 3.

Table 2: Example of the curriculum mapping process

US school grade	Major work and associated content	When covered in England?						
		Yr1	Yr2	Yr3	Yr4	Yr5	Yr6	Yr 7–9
K–2	**Addition & subtraction**	✓	✓	✓				
K	Join and separate sets of objects. (Writing of calculations encouraged but not required.)	[a]						
1	Add and subtract whole numbers within 20	✓						
1	Develop methods to add within 100		✓					
1	Develop methods to subtract multiples of 10		✓					

[a] This would be covered in the foundation curriculum.

Reporting findings from curriculum mappings

Mapping outputs: SMP and SMC overlap with the national curriculum

Table 3 and Table 4 show the outcomes of the analysis for the SMP and the SMC respectively. In Table 3 the symbols show the number of sentences within each SMP that were fully matched (a tick), partially matched (a tilde) or not matched (a cross) to the national curriculum for each of the key stages, as well as a summary across the three key stages. In Table 4, the ticks show when the SMC associated with each major work would be taught in the national curriculum. The shaded columns show the year groups and grades that are equivalent to each other.

Table 3: Comparison of the SMP to the national curriculum

Common Core standard	KS1	KS2	KS3	Across KS1–3
1. Make sense of problems and persevere in solving them	✓✓ ~ ✗✗✗✗✗✗✗	✓✓ ~~~ ✗✗✗✗✗	✓✓✓ ✗✗✗✗✗✗✗	✓✓✓✓✓ ~~ ✗✗✗
2. Reason abstractly and quantitatively	✗✗✗	✗✗✗	~ ✗✗	~ ✗✗
3. Construct viable arguments and critique the reasoning of others	✓✓ ✗✗✗✗✗✗✗	✓✓✓ ✗✗✗✗✗✗	✓✓✓ ✗✗✗✗✗	✓✓✓✓ ✗✗✗✗✗
4. Model with mathematics	✓✓ ✗✗✗✗✗	✓✓✓✓✓ ~ ✗	✓✓✓✓✓ ~ ✗	✓✓✓✓✓ ~
5. Use appropriate tools strategically	✗✗✗✗✗✗	✓ ✗✗✗✗✗	✓ ~ ✗✗✗✗	✓ ~ ✗✗✗✗
6. Attend to precision	✗✗✗✗✗✗✗	✓ ✗✗✗✗✗	✓ ✗✗✗✗✗✗	✓✓ ✗✗✗✗✗
7. Look for and make use of structure	✓ ✗✗✗✗✗	✓ ✗✗✗✗✗	✓✓ ✗✗✗✗	✓✓ ✗✗✗✗
8. Look for and express regularity in repeated reasoning	✗✗✗✗✗✗	✗✗✗✗✗✗	✗✗✗✗✗✗	✗✗✗✗✗✗

Table 4: Comparison of the SMC to the national curriculum

US grade	Major work for grade(s)	When covered in England?						
		Yr1	Yr2	Yr3	Yr4	Yr5	Yr6	Yr7–9
K–2	Addition & subtraction	✓	✓	✓				
3–5	Multiplication & division of whole numbers	✓	✓	✓	✓	✓	✓	
3–5	Multiplication & division of fractions	✓	✓	✓	✓	✓	✓	
6	Ratios & proportional relationships					✓	✓	✓
7	Arithmetic of rational numbers							✓
8	Linear algebra & functions							✓

Making comparisons from curriculum maps

Completed curriculum maps only form the first stage of a comparability study. The next stage requires the maps to be interpreted to compare the jurisdictions/curricula. Curriculum maps such as these can be used to identify similarities and differences in the content coverage, the ordering and progression of content across grades/school years, and the breadth and depth of the curriculum. In this section we will discuss these four comparisons and whether they could be made from our mappings given the approaches we used for the SMP (where age was only available for the national curriculum) and the SMC (where we used a sub-set of the content).

Content coverage

The most basic comparison that can be made is whether content from the master curriculum/jurisdiction is included within the other curricula/jurisdictions being compared (the comparators). These comparisons could be made for both the SMP (see Table 3) and the SMC (see Table 4). Our analysis of the SMC showed that almost all the content we mapped is included in the national curriculum – there were only four phrases without matching content. In contrast, there were considerable differences for the SMP, where half of the sentences could not be matched to the national curriculum. However, there were close matches for some of the individual SMP. Every sentence within the fourth SMP, "model with mathematics", could be matched to the national curriculum, with all but one of those being a complete match. Other SMP had good numbers of matches once the partial matches were included. For example, the first SMP, "make sense of problems and persevere in solving them" had complete matches for half its ten sentences and partial matches for a further two.

It is relatively easy to identify and code partially matched content; however, some content is not explicitly included but must be taught as other content relies on it, and this can be trickier to record. We found examples of implicit content in both the SMP and the SMC mappings. For example, the first of the SMPs requires students to explain the meaning of a problem and find entry points to a solution. We could not find refences to this in the national curriculum, but it does require students to solve problems and they cannot do this without working out what the problem means and trying to find an entry point to solve it. Therefore, we noted it as an implicit match.

We also found examples where the content from the CCSS was not included in the national curriculum content, but it was mentioned in the accompanying non-statutory notes and guidance. For example, the fractions content of the CCSS expects students to “Explain why procedures for multiplying fractions make sense” (NGA Center & CCSSO, 2010, p. 33). This is not included in the national curriculum content for multiplying fractions, but students would need to know this to be able to confidently multiply fractions (year 6 content). In addition, the non-statutory guidance states that “pupils should use a variety of images to support their understanding of multiplication with fractions” (Department for Education, 2013b, p. 41), which is similar. Therefore, we coded it as an implicit requirement.

Placement of content in grades/school years

When the content within curricula is allocated to particular school grades or year groups, it is possible to compare the ages at which particular areas of content are introduced and how many years they are taught for. Both the SMC and the national curriculum allocate the content in this way, so we were able to make these comparisons of content. For example, the fractions mapping (see Table 5) showed that students in England begin to recognise and generate equivalent fractions at a much earlier age and are taught this content for more years than students in the US, where this is only a requirement in grade 4. However, students in both countries learn to multiply fractions by whole numbers at the same age. Such comparisons were not possible for the SMP, as these standards are common to all grades in the US.

Table 5: Extract from the curriculum mapping of the SMC – multiplication and division of fractions

US school year group	Major work and associated content	When covered in England?						
		Yr1	Yr2	Yr3	Yr4	Yr5	Yr6	Yr7 –9
4	Recognize & generate equivalent fractions		✓	✓	✓	✓		
4	Compose/decompose fractions from/into unit fractions							
4	Multiply a fraction by a whole number					✓		
5	Add and subtract fractions with unlike denominators					✓		

Depth of the curriculum

Curriculum maps can be used to compare the depth of the content coverage that students are expected to learn. Although depth can refer to the difficulty of the knowledge that students have learned in a particular area, it is more often used to indicate the amount of knowledge they have gained in that area within a period of time.

Regardless of which definition of depth is used, comparisons of the depth of curriculum are more difficult than considering whether content is present and when it is taught because a degree of expert judgement is required in order to consider whether additional and omitted content balance each other out. For example, Table 5 shows that there are differences in the fractions content included in grade 4 of the SMC and year 5 in the national curriculum. Anyone making a comparison of depth would have to consider how the additional content on adding and subtracting fractions in the year 5 curriculum compared to omission of the content on composing and decomposing from/into unit fractions and starting the equivalent fractions content in earlier year groups.

Despite the difficulties in making these judgements, and the requirement for expert opinion in order to make accurate judgements, it is possible to make some comparisons of the depth of content for both mapping methods. For the SMC mappings, which focused upon particular content areas, experts could look at those mappings and use them to decide whether students would have acquired a greater depth of knowledge in that area during a particular school grade/year, or whether they had acquired more difficult knowledge before a certain point in their schooling. However, they could not make an overall judgement about the depth of knowledge that was taught in a particular grade/year across all areas of mathematics.

The SMP mapping only allows comparisons of the depth of knowledge acquired over the course of schooling, as the content is common to all year groups. For example, there is almost complete overlap in the coverage of the fourth SMP, “Model with mathematics”, so students are likely to achieve the same depth of knowledge in this area. In contrast, the final SMP, “Look for and express regularity in repeated reasoning”, appears to be entirely absent from the national curriculum, so we can be reasonably confident in stating that students following the CCSS would have acquired a greater depth of knowledge in that area.

Breadth of the curriculum

Curriculum maps can also be used to make comparisons about the breadth of the curriculum coverage, either within particular grades/school years or across the whole of the curricula being compared. In order to make comparisons about the breadth of the curriculum coverage, it is necessary to map all the content from each curriculum that is used in the study. This means that as well as mapping matching content from all the comparator curricula to the master curriculum, it is necessary to record the content within each of the comparator curricula that is not included in the master curriculum. As comparing the entire content was out of the scope of our study, it was not possible to identify the breadth of either curriculum from our mappings.

Affordances and limitations of the methods for curriculum mapping and the resulting comparability claims

In this article we have described three different methods of curriculum mapping: (1) mapping the entire content, (2) mapping selected content and (3) mapping curricula structured differently by age. We have also considered the different sorts of comparisons that can be made from curriculum maps – content coverage, when taught, depth and breadth of coverage – and have discussed which comparisons can be made for each approach (see Table 6). In this section, we will consider the affordances and limitations of the different approaches to curriculum mappings and the comparability claims that can be made from them in these three approaches.

Table 6: Summary of the comparisons that can be made from each method

	Entire content	Different age structures	Selected content
Content coverage	Yes	Yes	Yes**
Content placement in grades/years	Yes	No	Yes**
Depth of curriculum	Yes	Possibly*	Partially**
Breadth of curriculum	Yes	Possibly*	No

* possible across multiple grades/years if the start and end grades/years align
** only possible for selected content areas

Generally (all methods)

Curriculum mapping is a very useful method for identifying differences in what is taught in terms of the content that is covered and the year in which it is taught. For example, our mapping of the SMC showed that, for the areas we looked at, there is very little difference in the content that is included in the SMC and the national curriculum, but the content is generally introduced earlier and taught over a greater number of years in the national curriculum. The visual nature of the mapping documents enables a focused comparison of the curricula (Greatorex et al., 2019) and allows comparisons to be made with relative ease (Elliott, 2014). These comparisons can be used to see what is happening at a particular time (Elliott, 2014) or to study differences between current and older versions of curricula (Greatorex et al., 2019). The maps may also provide insights into the approaches to a subject in the two countries. For example, while carrying out our mapping, we were able to identify that the CCSS had an emphasis on conceptual knowledge as well as procedural knowledge, whereas the national curriculum emphasised procedural knowledge.

However, there are limitations that should be considered. The mapping document enables the comparisons between curricula rather than providing instant answers about the comparability of curricula. Curriculum maps should be interpreted by subject experts (Elliott, 2014) who may then go on to make comparability claims. Summaries of the experts' interpretations are often given greater prominence in the resulting reports than the curriculum maps that they are based upon. The requirement to summarise the maps can introduce errors into the analysis, particularly where the interpreter is more familiar with the content in one curriculum than the others. Other misinterpretations could be introduced when terminology is used differently within the curricula meaning that identical content goes unmatched, or content is matched incorrectly. Finally, curriculum mapping can only provide information about the intended curriculum; it cannot provide insights into what is taught in schools or how it is taught (the enacted curriculum).

Mapping the whole curriculum

The most comprehensive mapping that is possible is when the whole comparator curriculum is compared to the whole of the master curricula. We saw in the previous section that this enables all four types of comparisons to be made – content coverage and placement, and the breadth and depth of the curricula. If content from multiple years is mapped, it is also possible to compare the progression in understanding across grades/years.

Bearing this in mind, it may seem difficult to justify moving away from this approach; however, there are some disadvantages to mapping entire curricula. Firstly, mapping is a time-consuming exercise. The greater the quantity of content that is mapped, the longer it takes and the more it costs. A second consideration is the amount of information that is produced and the usefulness of that information given the aims/purposes of the comparability study. In order to make useful observations and interpretations regarding mapping claims it is necessary to use the mapping to make one or more of the four types of comparisons. The more content that is mapped, the more difficult it is to make these comparisons. Even identifying similarities and differences between curricula can prove difficult when there are many pages of a mapping document to consult. Similarly, although it is possible to make comparisons of the depth and breadth of the curricula, it may be very difficult for an expert to decide how the multiple differences in the breadth and depth of coverage in each area of the curriculum balance out, and therefore to draw conclusions about which curriculum contains that greatest depth or breadth of content. A final limitation is that it does not provide any information about the importance of particular areas of content.

Mapping limited content

Including only certain topics, as we did for the SMC, is a pragmatic approach that still enables most types of comparisons to be undertaken. It may also make it easier to identify similarities and differences across the curricula being compared as there is less data to consider. However, this approach is inevitably less robust than comparing whole curricula as there is no information about the omitted content. The omission of content also precludes comparisons and claims about the depth or quantity of the content included in particular grades/years, as it is unlikely that the quantity of content contained within the selected areas is representative of the quantity of content in the omitted areas. Taking our mapping of the SMC as an example, we found that more areas of mathematics were included in the national curriculum for pupils in years 1–3 than were included in grades K–2 of the SMC. However, it would not be appropriate to use this finding to claim that the national curriculum contained a greater depth of content as it does not take into consideration the content areas such as geometry that were excluded from our mapping.

Perhaps the greatest difficulty with this approach is selecting content for the mapping that will enable meaningful comparisons to be made. This could be areas of the subject that have been identified as particularly important (e.g., the major works associated with the CCSS that we used in our mapping), but it could also be one or more domains within a subject (e.g., number as a domain of mathematics) or particular areas within a domain (e.g., fractions as an area within number). Whatever domain or area is chosen, it is important that there is an underlying justification for the choice. This will help to ensure that the resulting claims of comparison are useful and will reduce the likelihood of self-fulfilling claims resulting from the careful selection (or deselection) of content.

Mapping curricula with different age structures

These comparisons are effectively a subset of the whole curriculum mappings, but where one curriculum is arranged differently from another. One may have separate content for every age group (like the KS1 & 2 national curriculum in England) when its comparator curricula combine several year groups together or have identical content for all age groups (like the SMP). The researcher will not have any choice about whether to use this approach, as it is a characteristic of the documents they are working with. This was the case with our SMP mapping, which showed that it was still possible to make meaningful comparisons when working with curricula with this issue. This approach to mapping shares the affordances of mapping whole curricula, for example allowing most types of comparison to be undertaken and allowing comparisons of the depth and breadth of content covered over the whole of the age range. We were able to identify standards within the SMP that had different depths of content to the national curriculum, such as the 5th standard which contains requirements to use technological tools that had no equivalent in the national curriculum. When the mappings for all eight SMP are considered, there appears to be greater depth in the SMP content than in the national curriculum.

This approach also suffers from the same limitations as mapping the whole curricula. Moreover, it is more limited than other mappings of whole curricula in that it cannot be used to explore differences in the age at which particular topics are taught, or in the amount of content for particular age groups. Thus, although we identified areas of the SMP that are also covered by the national curriculum, such as “Model with mathematics”, we could not state whether the CCSS require students to learn more content or to have greater knowledge of the content in a particular grade than would be expected in the equivalent year of the national curriculum.

Conclusion

Within comparability, curriculum mapping is used to analyse similarities and differences in the content of multiple curricula. It is important to note that it only provides insights into the intended curriculum; it cannot provide information about the taught or learned curricula, or the teaching methods that are adopted in classrooms. Although the preferred approach is for whole curricula to be compared, there will be occasions where this is not possible due to time constraints, lack of funding, or where the researchers are only interested in part of the curriculum. Our study has shown that it is also possible to use this method to map a sub-set of the content and make meaningful comparisons and claims from the mapping, provided that the content has been selected in a way that can be justified. Thus, we can use our mapping to claim that the number and algebra content contained within the SMC and the national curriculum is comparable.

However, it would not be appropriate to extend the claim to say that the national curriculum and CCSS are comparable for the whole mathematics curriculum, nor could we infer the comparability of other areas of mathematics on the basis of the areas we mapped.

There can be issues with curriculum mappings where the curricula that are used in the comparisons are structured differently. Some structural differences, such as the content appearing under different headings, may not affect the mapping or the comparisons that can be made from it. Other differences, such as differences in the way in which the content is structured by age, can affect the comparisons by restricting what can be compared or the precision of those comparisons. We showed that it was still possible to map the content when there were differences in the age structures of the documents, and to make justifiable claims on the basis of the mapping, but we could only do this for the whole document rather than for individual year groups. In the case of the SMP mapping, we can claim that the SMP require students to have greater understanding of mathematical processes than the national curriculum, but we could not state whether this was true for students in particular grades/years. We also could not claim that students in the US would be better at these skills than students in England, as students may be taught skills that are not included within the curriculum.

Both approaches we used (mapping a sub-set of content and mapping curricula that are structured differently by age) enabled us to make claims of comparisons of the similarities and differences in the content that is included and the depth of the content that is taught; however, the approaches did limit the other comparisons that were possible. When a sub-set of the content was mapped, it was not possible to compare the breadth of the content. Therefore, we cannot use our SMC mapping to compare the breath of the national curriculum to the breadth of the CCSS. When the content within one or more of the curricula was common to multiple age groups, it was not possible to compare the age when the content was taught. Therefore, we cannot make claims about the skills taught to equivalent year groups, or that students of a particular age would be expected to demonstrate.

This article has introduced the types of comparisons that can be made from curriculum mapping studies generally, and when features of the curricula or the study design affect the mapping that can be carried out. However, the approach that is chosen will affect the claims that can be made about the comparability of different curricula. Therefore, any researcher wishing to use curriculum mapping as the basis for a comparability study must balance the intentions of the comparability investigation with the rigour of the methodological approach that they use.

Future research may want to consider how the selection of content that is mapped can affect the claims that can be made, and how the comparisons that can be made from mappings are affected when the curricula are for skills-based subjects, such as English literature or foreign languages, rather than content-based subjects like mathematics.

References

Achieve the Core. (n.d.). *CCSS Where to focus Kindergarten mathematics*.

Alcántara, A. (2016). *International Baccalaureate mathematics comparability study: Curriculum and assessment comparison*. International Baccalaureate Organization.

Common Core State Standards Initiative. (2022). *About the Standards*.

Department for Education. (2012). *Review of the National Curriculum in England: What can we learn from the English, mathematics and science curricula of high-performing jurisdictions?*

Department for Education. (2013a). *Mathematics programmes of study: key stage 3.*

Department for Education. (2013b). *Mathematics programmes of study: key stages 1 and 2.*

Elliott, G. (2014). Method in our madness? The advantages and limitations of mapping other jurisdictions' educational policy. *Research Matters: a Cambridge Assessment publication, 17*, 24–28.

Greatorex, J., Rushton, N., Coleman, T., Darlington, E., & Elliott, G. (2019). *Towards a method for comparing curricula*. Cambridge Assessment.

Kridel, C. A. (2010). *Intended curriculum. In Encyclopedia of curriculum studies* (Vol. 1, pp. 179–181). Sage Publications.

NGA Center, & CCSSO. (n.d.). Key shifts in mathematics.

NGA Center, & CCSSO. (2010). *Common Core State Standards (Mathematics)*. National Governors Association Center for Best Practices & Council of Chief State School Officers.

Ofqual. (2012). *International comparisons in senior secondary assessment full report: Table supplement*. Ofqual.

Schmidt, W. H., Houang, R. T., Cogan, L. S., & Solorio, M. L. (2018). The 1995 TIMSS Curriculum Analysis and Beyond. In *Schooling Across the Globe: What We Have Learned from 60 Years of Mathematics and Science International Assessments*. Educational and Psychological Testing in a Global Context. Cambridge University Press, 43–180.

UK Government. (n.d.). *The national curriculum.*

Exploring speededness in pre-reform GCSEs (2009 to 2016)

Emma Walland (Research Division)

Background literature and research aim

The speededness of an assessment refers to the extent to which the assessment's time allocation influences students' performance, or the extent to which the assessment occurs under time pressure (Schnipke & Scrams, 1997). Speededness could be considered a source of increased demand[1] in an examination, along with many other sources that have already been explored (Ahmed & Pollitt, 2000; Crisp et al., 2008; Fisher-Hoch et al., 1997; Pollitt et al., 2008). Fisher-Hoch and Hughes (1996) proposed that demand can be valid or invalid. Valid demand is intended by the setter and related to the constructs being assessed, whereas invalid demand is unintentional, and can arise for several reasons. Speededness can negatively affect students' experiences of taking an assessment and, therefore, insufficient time allocation could be considered a source of invalid demand in an assessment that is not intended to be speeded. Skilled setters use their experience to determine proper assessment length; however, this can be a challenging task. The number and nature of items may depend on the age of students, the time available for testing, the type of items used and the type of interpretation to be made (Directorate for Quality and Standards in Education, Malta, 2022).

One way to estimate intended speededness is through the number of marks per minute of the assessment. The higher the number of marks per minute, the more speeded the assessment is likely to be. Whether the students experienced an assessment as speeded can also be explored by analysing not-reached items, or items left blank at the end of students' examination papers. Other methods look at response data to determine a point in the assessment where student performance deteriorated, as an indication of potential speededness (Shao et al., 2016). However, an important caveat is the use of "ramping" in some examination designs, whereby easier items are put at the start of the paper and the more difficult ones occur later. This means that items left unanswered at the end (or where student performance deteriorated) could also be due to students finding them too demanding. In England, conventional wisdom and research evidence indicate that GCSE students answer items sequentially (Spalding, 2011b) and GCSE examination developers across a range of subjects make use of ramping across

[1] I use the term "demand" to refer to how challenging students find their examinations.

the examination paper (Johnson et al., 2017; Johnson & Rushton, 2019; Spalding, 2011a).[2] There is also evidence of ramping within each main question, with ramping occurring in each item that is part of an overall question (Johnson & Rushton, 2019). This must be considered when examining the items left unanswered at the end of a student's examination paper. Thus, speededness of an examination would be more strongly evidenced by higher ability students (rather than lower ability) leaving items blank at the end of a paper. The relationship for lower ability students can be complex. Pohl et al. (2014) argued that certain assessments that are highly speeded (e.g., reading tests) might have higher levels of omitting at the end for higher ability students than lower ability ones due to their different test-taking strategies. This could be because higher ability students may have worked more carefully on getting the items right whereas lower ability students may have skipped through the assessment quickly due to not being able to answer many of the items. They argue that the same pattern may not appear in less speeded assessments.

Other factors that could influence omit rates at the end are student motivation (Matters & Burnett, 2003; Pohl et al., 2014) and guessing. The former should not be a major concern for high-stakes examinations such as GCSEs. Regarding the latter, methods to detect guessing have been developed for multiple-choice assessments, where students are likely to engage in rapid non-systematic guessing (Schnipke & Scrams, 1997; Yamamoto, 1995), but these are not suitable for less constrained items where students are less likely or less able to guess and detecting guessing would be much more complex (Jones, 2019; Pollitt et al., 2008). Lastly, students' test-taking behaviour may be influenced by their personal characteristics. Matters & Burnett (2003) found that test-irrelevant thinking and academic self-concept predicted whether students were likely to omit short-response test items.

While GCSE written examinations are not intended to be speeded, there has been little research exploring this (as noted by Spalding (2011a) and Wheadon (2011)). Further investigation of this is important for assessing the fairness and validity of assessments. The aim of this research was to investigate the speededness of past GCSE written examinations, using a method that only considered the scored responses to items and whether they were omitted.

Method

Data

I selected a sample of 340 GCSE written examination components[3] for analysis. These components were from Physics, Science, Chemistry, Biology and Mathematics qualifications offered by OCR. These components were anticipated to have large entries as well as large numbers of items. None of the components in the sample had optional items. The main reason for that was that such items would show as missing in the data and, therefore, would confound the analysis.

[2] Note that ramping is not suitable for all types of examinations, for example, English Language GCE, which only has a few long answer items intended to be of equal demand.

[3] GCSEs are made up of separate exams or non-exam assessments called "components".

The time period for analysis was from 2009 to 2016, prior to the GCSE reform (Ofqual, 2018).[4] I included both higher and foundation tier components,[5] and excluded components with fewer than 100 entries from the analysis.

I conducted the analyses using SAS Enterprise Guide (version 7.1). For each component the following information was available: number of items, maximum mark for each item, and time allowed. For each student the data consisted of the scored response on each item, or an indicator of "missing" if the student had not attempted the item.

Prior to analysis I computed indicators of potential speededness for each component. These consisted of the marks per minute and the average (mean) percentage marks lost from the longest string of unanswered items at the end of each student's examination paper, referred to as "average percentage lost (at the end)". For each component, I also calculated: 1) the average percentage lost (at the end) per quartile of student achievement; and 2) the median percentage marks lost. The final dataset included all the above component-level data and was used for the main analyses.

Data analysis

First, I analysed the dataset descriptively in terms of marks per minute and the average percentage lost (at the end), and how this differed for different achievement quartiles and tiers. This enabled me to identify potentially speeded components. I only considered a component to be potentially speeded if the average percentage lost (at the end) was high for higher achieving students in addition to, or instead of, lower achieving students. Lower achieving students taking a highly speeded paper may not have a high percentage of marks lost at the end of their papers. This is because, theoretically, they could progress more quickly through their papers and reach the end due to not being able to answer many of the items earlier on in the paper (Pohl et al., 2014).[6]

Results and discussion

Of the total 340 components I analysed, there were 78 Mathematics, 78 Physics, 72 Science, 68 Chemistry and 44 Biology components. 170 of them were foundation and 170 were higher tier. Years ranged from 2009 to 2016, with 2012 being the most represented year. Table 1 shows the descriptive statistics for the numeric variables, rounded to two decimal places unless otherwise stated. Table 1 shows that the number of students in each component ranged from 137 to 55 564, with a mean of 14 185 (rounded to the nearest whole number). Examination duration ranged from 40 minutes to two hours, with a mean of just over an hour. The number of items per component ranged from 19 to 66, with a mean of 34 items.

[4] The reform led to a new grading scale being used, among other changes.

[5] Tiering is used in some GCSEs (e.g., Science and Mathematics) to better allow for the wide range of abilities at this level. For the assessments in this analysis, foundation tier components were graded C to U, and higher tier components were graded A* to E.

[6] Student motivation can also be a factor influencing test completion. However, in this context, I assumed that student motivation was generally high as GCSEs are high-stakes examinations for students.

Table 1: Descriptive statistics for numerical variables (n=340 components)

Variable	Median	Mean	Std Dev	Minimum	Maximum
Number of students	11 666.50	14 185.18	13 295.70	137.00	55 564.00
Exam duration (minutes)	60.00	65.47	21.81	40.00	120.00
Total number of items	30.00	34.41	11.08	19.00	66.00
Maximum raw mark	60.00	64.63	18.69	42.00	100.00
Marks per minute	1.00	1.00	0.06	0.83	1.11
Average (mean) % lost (at the end)	0.65	0.81	0.65	0.01	4.05
Median % lost (at the end)	0.00	0.01	0.18	0.00	3.33
Average (mean) % lost (at the end) (Q0)	1.95	2.29	1.73	0.04	8.08
Average (mean) % lost (at the end) (Q1)	0.37	0.58	0.59	0.00	3.90
Average (mean) % lost (at the end) (Q2)	0.18	0.27	0.33	0.00	3.49
Average (mean) % lost (at the end) (Q3)	0.05	0.10	0.22	0.00	3.34
Average (mean) % completed	92.97	90.80	8.92	0.06	99.93
Average (mean) % not completed	7.03	9.20	8.92	0.07	99.94
Median % lost (at the end) (Q0)	0.00	0.09	0.55	0.00	5.00
Median % lost (at the end) (Q1)	0.00	0.02	0.24	0.00	3.33
Median % lost (at the end) (Q2)	0.00	0.01	0.18	0.00	3.33
Median % lost (at the end) (Q3)	0.00	0.01	0.18	0.00	3.33

Note. Q stands for Quartile, with 0 representing the lowest achieving quartile and 3 the highest. Average % completed and not completed refers to the percentage of students who completed or failed to complete their examinations, respectively. Where the term "average" is used, it refers to the mean.

Marks per minute

Table 1 shows that the mean marks per minute across all components was approximately 1, ranging from 0.83 to 1.11. This does not appear to be problematic in terms of speededness. For example, online guidance given to students in the context of GCSE History suggests that 1 mark per minute is a good rough guide to work towards (OCR, 2024). However, this does also depend on the nature of the items, for example, how much reading the item requires and how long it takes to produce a response. The distribution of marks per minute for each subject group is illustrated in Figure 1. The figure shows that most of the marks per minute across subject groups were around 1. There was more of a range in the data for Mathematics than for the other subjects.

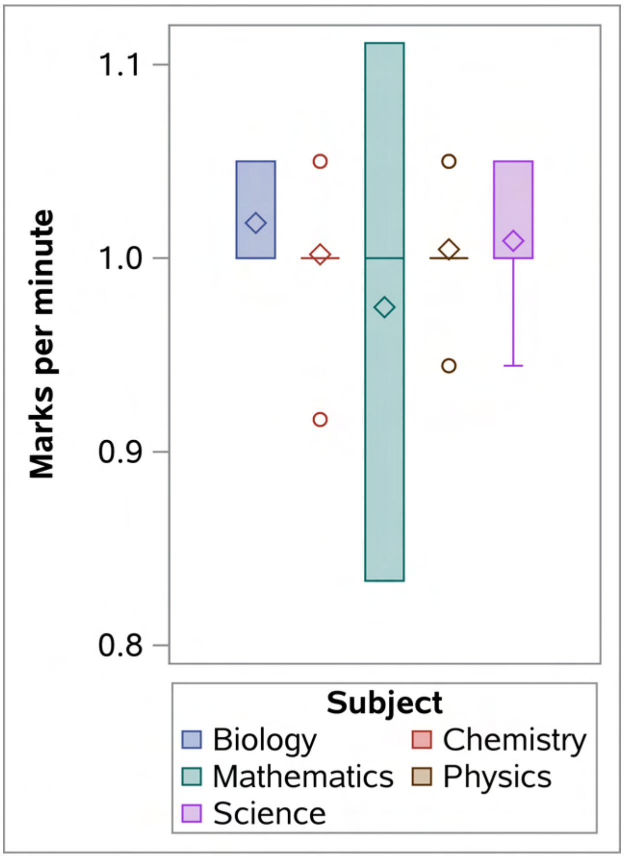


Figure 1: Schematic box and whisker plot showing the distribution of marks per minute for each subject group for all components. The diamonds represent the means, the box height represents the interquartile range, the horizontal line in each box represents the medians and the circles represent outliers. (The two subjects with no visible boxes had nearly the entire distribution clustered near 1 with only a few outliers).

There were 21 components with the highest number of marks per minute (1.1) and they were all Mathematics components. Marks per minute are about the **intended** speededness of the assessment, as it is determined at the design phase. The results of the analyses here indicate that the exams were likely not intended to be speeded based on the marks per minute. The average percentage lost (at the end), however, can indicate how the students **experienced** the assessment and whether it may have been experienced as speeded. I explore this subsequently.

Average percentage lost (at the end)

As shown in Table 1, the average percentage marks lost (at the end) by students across all achievement groups and tiers was 0.81 per cent (SD=0.65) and the median (of the medians across all achievement groups) was 0 with a range of 0 to 3.33 per cent. Across all components, an average of 90.80 per cent of students completed their examinations. This indicates that overall, there was little average percentage lost (at the end) across the sample. Figure 2 shows that the average percentage lost (at the end) was slightly higher for Mathematics components, and was very similar for the three single sciences (Biology, Chemistry and Physics).

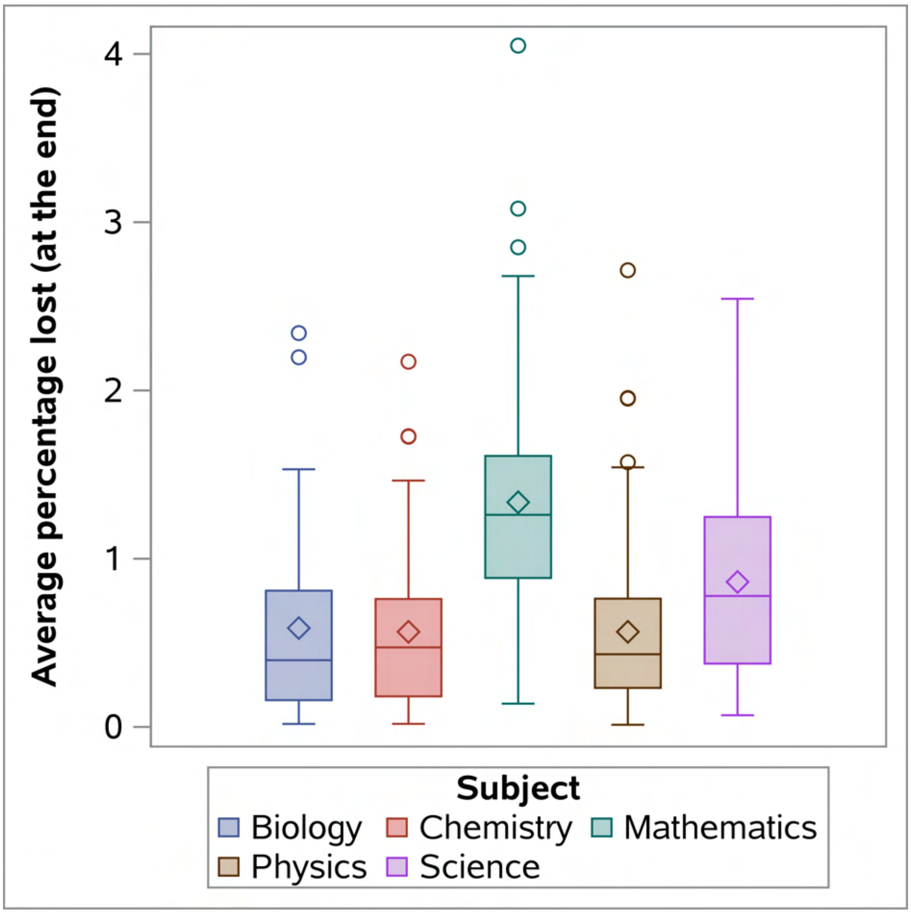


Figure 2: Schematic box and whisker plot showing the average percentage lost (at the end) for each subject group included in the analysis. The diamonds in the boxes represent the means, the horizontal lines within each box represent the medians, and the box height represents the interquartile range. The circles outside the boxes represent outliers.

I examined next the average percentage lost (at the end) for different student achievements and different tiers. The average percentage lost (at the end) for the highest achieving students in each component was much smaller, at 0.10 per cent. For the lowest achieving students it was much larger than for all students, at 2.29 per cent (Table 1). As noted, if a component was speeded, we would expect to see high levels of average percentage lost (at the end) for higher achieving students. Foundation tier components had higher levels of marks lost (at the end) in general compared with their higher tier counterparts, as shown in Figure 3. The difference between foundation and higher tier components decreased as student achievement increased (i.e., as we move from Quartile 0 to Quartile 3).

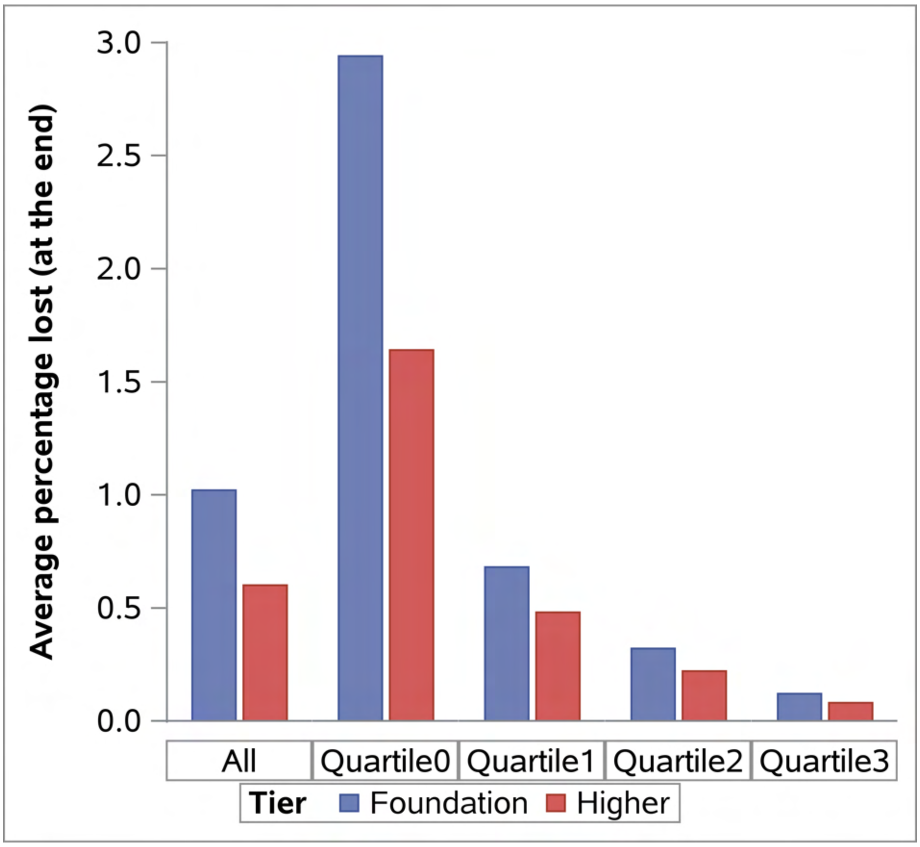


Figure 3: Average percentage lost (at the end) for each tier and each quartile of achievement, and overall

If an assessment was speeded, and students experienced time pressure, there is likely to be strings of items missing at the end for higher achieving students (Pohl et al., 2014). Given that there were few instances of omission at the end for higher achieving students, this suggests that most GCSEs were not speeded and items missing at the end were more likely due to student ability. Regarding low ability students, they likely experienced the items at the end as demanding and omitted them. But there are other possibilities: for example, that lower ability students are slower workers in general or have lower levels of motivation to complete their assessments. As noted previously, one theory according to Pohl et al. (2014), is that in a highly speeded assessment, lower ability students would omit fewer items at the end than higher ability students due to differences in test-taking strategies. I explored the data to investigate any instances where this pattern occurred, and no such examples were found.

The average percentage lost (at the end) for all components, for all students together (on the left) and for the highest achieving students (on the right), is illustrated in Figure 4. This shows that most of the components had a very low average percentage lost (at the end). However, there were some outliers with relatively high values, which could indicate speededness.

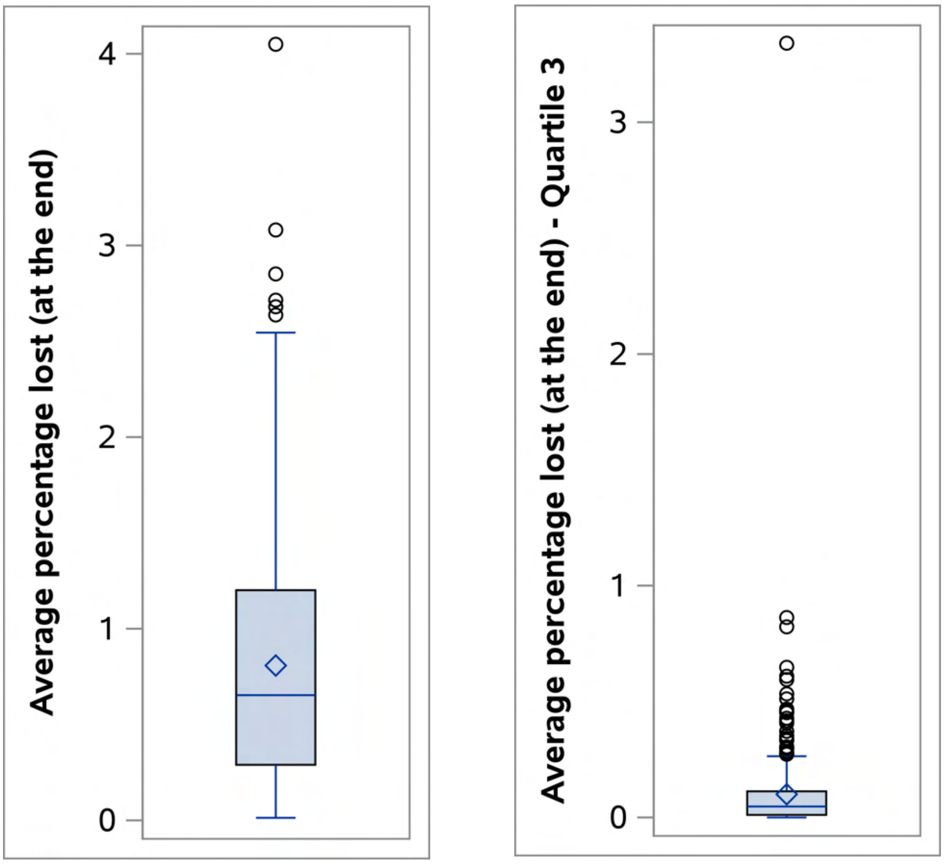


Figure 4: Schematic box and whisker plots showing the average percentage lost (at the end) for all components in the analysis. The figure on the left is the overall average percentage lost (at the end) across all students and the figure on the right is for the highest achieving quartile. The diamond represents the group mean, the box height represents the interquartile range, and the horizontal line within the box represents the median. The circles represent the outliers.

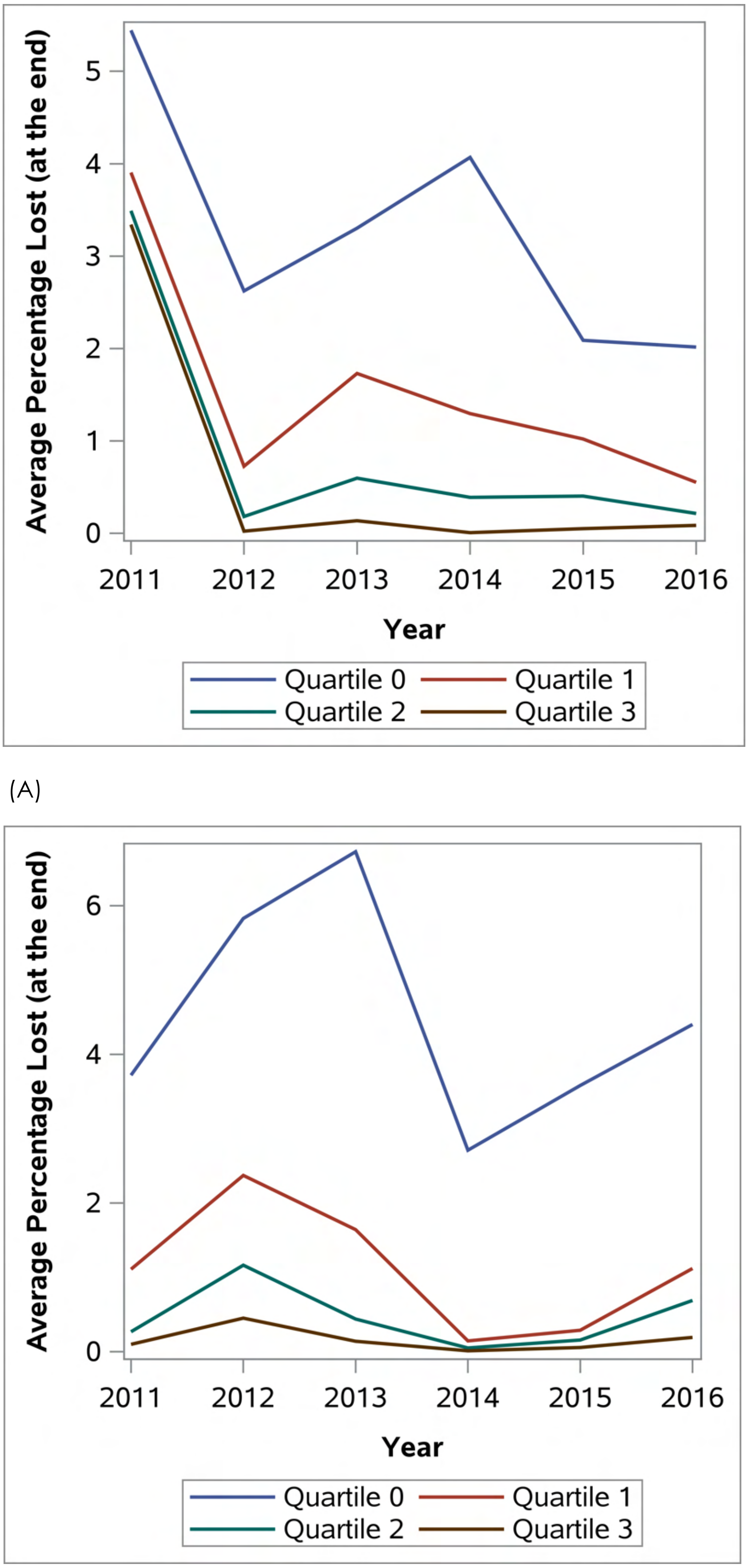
I looked at the components with the highest average percentage lost (at the end) for all students (Appendix A), as well as for higher achieving students (Appendix B). When all students were considered, there were 17 components which had an average percentage lost (at the end) above 2.00 per cent. Of these, the top six would be considered outliers according to Tukey's fences[7] (the upper bound was 2.57 per cent). For the students in the highest achieving quartile (Q3), 29 components were identified as outliers using Tukey's fences (in this case the upper bound was 0.26 per cent).

[7] Tukey's fence is a method to detect outliers. Outliers are defined as values higher than Q^3 + 1.5(IQR) and lower than Q^1 – 1.5(IQR). Q^3 refers to Quartile 3, Q^1 to Quartile 1 and IQR to the inter-quartile range (Q^3-Q^1).

Analysis of the component with the highest average percentage lost (at the end)

The component with the highest average percentage lost (at the end) was the Mathematics M101 higher tier paper in 2011. The marks per minute was 1, which was the same as for the equivalent component in other years. This component also had the highest average percentage lost (at the end) for higher achieving students, at 3.34 per cent. In fact 99.94 per cent of the students who took this examination (11 527 students) did not complete it. However, nearly all these students did reach the penultimate item. The median last item attempted was the penultimate one and the median number of marks lost was 2, which was the tariff for the last item. As the exam was out of 60 marks, the last item was worth 3.33 per cent of the assessment. This suggests that students either ran out of time to reach the last item (slight speededness), or the last item was particularly challenging, even for higher achieving students, and so they omitted it. As the entry was large, the finding is unlikely to be related to the particular cohort.

I present the data for the Mathematics M101 component in the following graphs. Figure 5A shows the data for the higher tier and Figure 5B for foundation tier. Figure 5A shows that the average percentage lost (at the end) was higher in 2012 than in the other years, for all but the highest achieving quartile. Figure 5 shows that the percentage lost was highest for lower achieving students in all years and in both tiers. The number of students sitting this exam (across both tiers) ranged from 1439 to 15 148.



(A)

(B)

Figure 5: Average percentage lost (at the end) for higher tier (A) and foundation tier (B) students for Mathematics (M101)

Figure 6 shows the omit rate as a function of item position for Mathematics M101 higher tier in 2011. The omit rate for an item refers to the proportion of students who had no recorded response for that item. The figure shows that most items were completed by most students throughout the assessment, but that the final item was not completed by most students.

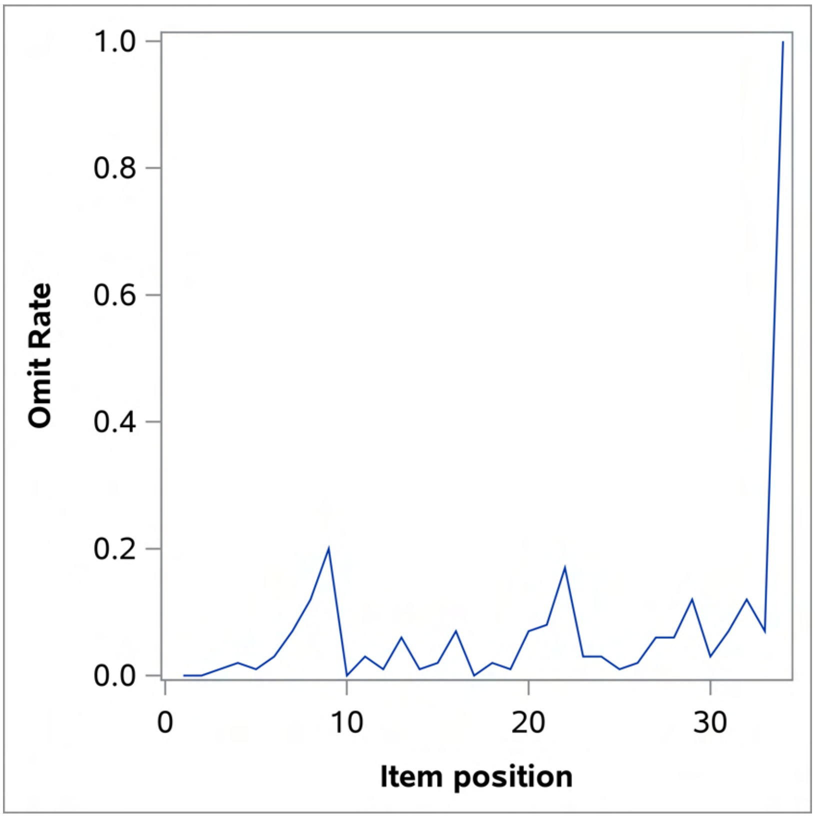


Figure 6: Omit rate as a function of item position, for Mathematics M101 higher tier in 2011

Figure 7 shows the facilities of each item, in the order they appear in the assessment. The facility of an item refers to the proportion of students who responded correctly to the item, which is generally used as an indicator of how easy an item is. The figure shows that the final item had a very low facility, which would usually indicate a demanding item (i.e., an item that a large proportion of students did not respond correctly to). However, as this statistic is based on the number of students who attempted the item but did not correctly answer it, it is not very informative in this case because only seven students attempted the item.

Figure 7 shows some evidence of ramping across the paper and within each item. The items later on in the test and the later items within each question were generally more difficult than the earlier ones.

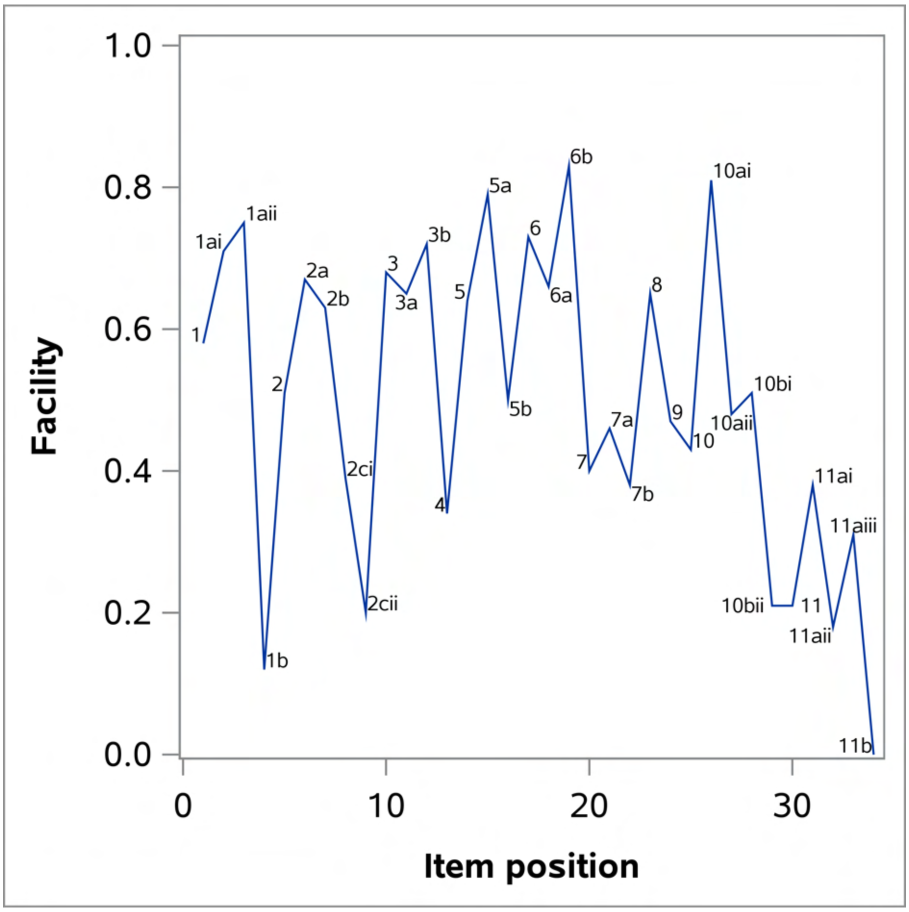


Figure 7: The relationship between item position and facility, for Mathematics M101 higher tier in 2011

Conclusion

The outcomes of this research show that there was little to no evidence of speededness in the pre-reform GCSE components analysed. Students did not seem to have been working under time pressure, according to the measure of speededness I used. This was indicated by very low levels of average percentage lost (at the end) for all, including higher achieving, students and also what appears to be appropriate marks per minute of around 1 mark per minute (OCR, 2024). This is reassuring as GCSEs are not intended to be speeded, and adds to our understanding of this under-researched area.

There were, however, some components, mainly from GCSE Mathematics, that showed some evidence of speededness. This could mean that students experienced time pressure, which could constitute an invalid source of demand, according to Fisher-Hoch and Hughes (1996). However, upon further investigation of the most outlying one it appeared that this was a result of most students not completing the last item. As there was only one item omitted, it is unclear whether students ran out of time or found it too demanding to attempt.

The limitations of using only data from scored responses (and omitted items in particular) to study potential speededness include that we are unable to account for student motivation and the role of test-taking strategies including guessing. The results are also complicated if there are higher tariff items at the end of an assessment, as students may start to write an answer but not complete it rather than omit it entirely. Speededness may also be indicated by students' performances deteriorating towards the end of a test, due to being rushed. Methods to detect this, such as change-point analysis (Shao et al., 2016), could be useful to complement the data on item omission. However, these methods may not function well in assessments with items that are ordered by increasing demand. The method I used is most useful in situations where motivation is high, items are not guessable, items are ordered by demand, students complete tests in order, there are no optional items, there are many items, and items each have small mark tariffs.

Bearing these limitations in mind, the research provides an example of a data-driven means of identifying assessments with potentially inappropriate time allowances using only data from scored responses. The method could be a useful tool to flag potentially problematic components which can then be investigated further. The data can be combined with other sources of data about speededness including post-administration surveys (see, for example, Steedle et al., 2022), and expert judgements about examination length. With the potential rise in computer-based tests, the data could also be used together with response time data in the future, to evaluate speededness.

Having methods to identify speededness can be useful for assessment designers and evaluators in relation to issues of validity and fairness relative to individual characteristics of students. Future research identifying whether particular students did not complete many of their examinations across different subjects and over time would also be interesting. This would lead to understanding whether running out of time is a stable personality trait that occurs across domains and over time, or whether it is specific to each assessment situation.

References

Ahmed, A., & Pollitt, A. (2000). *Observing context in action*. IAEA conference, Jerusalem.

Crisp, V., Sweiry, E., Ahmed, A., & Pollitt, A. (2008). Tales of the expected: The influence of students' expectations on question validity and implications for writing exam questions. *Educational Research, 50*(1), 95–115.

Directorate for Quality and Standards in Education, Malta. (2022). *Guidelines for paper setters: Educational assessment unit*.

Fisher-Hoch, H., & Hughes, S. (1996). *What makes mathematics exam questions difficult?* British Educational Research Association, University of Lancaster, England.

Fisher-Hoch, H., Hughes, S., & Bramley, T. (1997). *What makes GCSE examination questions difficult? Outcomes of manipulating difficulty of GCSE questions*. British Educational Research Association Annual Conference, York.

Johnson, M., Constantinou, F., & Crisp, V. (2017). How do question writers compose external examination questions? Question writing as a socio-cognitive process. *British Educational Research Journal, 43*(4), 700–719.

Johnson, M., & Rushton, N. (2019). A culture of question writing: Professional examination question writers' practices. *Educational Research, 61*(2), 197–213.

Jones, E. (2019). 'Multiple choice exams are easier than written exams!' *APMG International*

Matters, G., & Burnett, P. C. (2003). Psychological predictors of the propensity to omit short-response items on a high-stakes achievement test. *Educational and Psychological Measurement, 63*(2), 239–256.

OCR. (2024). *GCSE History A: How long should students spend on each question in the exams?*

Ofqual. (2018). *Get the facts: GCSE reform*.

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement, 74*(3), 423–452.

Pollitt, A., Ahmed, A., Baird, J., Tognolini, J., & Davidson, M. (2008). *Improving the quality of GCSE assessment.*

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*(3), 213-232.

Shao, C., Li, J., & Cheng, Y. (2016). Detection of test speededness using change-point analysis. *Psychometrika, 81*(4), 1118–1141.

Spalding, V. (2011a). *Is an exam paper greater than the sum of its parts? A literature review of question paper structure and presentation*. AQA.

Spalding, V. (2011b). *Structuring and formatting examination papers: Examiners' views of good practice*. AQA.

Steedle, J. T., Cho, Y. W., Wang, S., Arthur, A. M., & Li, D. (2022). Mode effects in college admissions testing and differential speededness as a possible explanation. *Educational Measurement: Issues and Practice, 41*(3), 14–25.

Wheadon, C. (2011). *An Item Response Theory approach to the maintenance of standards in public examinations in England* [Doctoral Thesis, Durham University].

Yamamoto, K. (1995). Estimating the effects of test length and test time on parameter estimation using the HYBRID model. *ETS Research Report Series, 1995(1)*, i-39.

Appendices

Appendix A – Components with the highest average percentage lost (at the end), all students

Code	Component name	Year	Average % lost	Average % lost Q0	Average % lost Q3	N students	Tier
M101	**Mathematics Unit B**	**2011**	**4.05**	**5.44**	**3.34**	**11 527**	**H**
M100	**Mathematics Unit A**	**2016**	**3.08**	**7.19**	**0.61**	**1635**	**F**
M103	**Mathematics Paper 1**	**2011**	**2.85**	**6.11**	**0.86**	**1194**	**F**
P100	**Physics A Modules P1, P2, P3**	**2012**	**2.71**	**7.20**	**0.21**	**3901**	**F**
M108	**Mathematics B**	**2013**	**2.68**	**7.48**	**0.11**	**17 103**	**H**
M100	**Mathematics Unit A**	**2011**	**2.64**	**5.94**	**0.51**	**6490**	**F**
S107	Science B: Unit 2 (B2, C2, P2)	2009	2.54	5.12	0.82	34 910	F
M101	Mathematics Unit B	2012	2.46	5.83	0.45	10 970	F
S109	Science Modules B2, C2, P2	2016	2.42	8.08	0.11	15 265	F
M108	Mathematics B	2013	2.40	6.46	0.10	17 155	H
S108	Science Modules B1, C1, P1	2012	2.39	5.92	0.65	14 203	F
B101	Biology A Modules B4, B5, B6	2012	2.34	7.84	0.37	1105	F
M108	Mathematics B	2014	2.29	5.22	0.42	30 484	F
B100	Biology A Modules B1, B2, B3	2012	2.20	5.99	0.46	3911	F
C105	Chemistry A: Unit 3	2010	2.17	6.81	0.15	634	F
M101	Mathematics Unit B	2013	2.17	6.72	0.14	1680	F
M107	Mathematics A	2009	2.06	5.28	0.17	22 963	F

Note: The six components that are outliers according to Tukey's upper fence are marked in bold. The assessment codes were created by the researcher and are not the real codes.

Appendix B – Components with the highest average percentage lost (at the end) for higher achieving students (Quartile 3)

Code	Component name	Year	Average % lost	Average % lost Q0	Average % lost Q3	N students	Tier
M101	Mathematics Unit B	2011	4.05	5.44	3.34	11 527	H
M103	Mathematics Paper 1	2011	2.85	6.11	0.86	1194	F
S107	Science B: Unit 2 (B2, C2, P2)	2009	2.54	5.12	0.82	34 910	F
S108	Science Modules B1, C1, P1	2012	2.39	5.92	0.65	14 203	F
M100	Mathematics Unit A	2016	3.08	7.19	0.61	1635	F
P100	Physics A Modules P1, P2, P3	2015	1.95	4.12	0.59	50 096	H
P107	Physics Modules P1, P2, P3	2012	1.46	3.58	0.53	843	F
M100	Mathematics Unit A	2011	2.64	5.94	0.51	6490	F
S105	Science A: Unit 3 (B3, C3, P3)	2012	1.90	4.49	0.47	6504	F
C100	Chemistry A Modules C1, C2, C3	2012	1.73	4.63	0.46	4194	F
B100	Biology A Modules B1, B2, B3	2012	2.20	5.99	0.46	3911	F
M101	Mathematics Unit B	2012	2.46	5.83	0.45	10 970	F
P104	Physics A: Unit 2 (P4, P5, P6)	2012	1.20	2.05	0.43	12 009	H
M108	Mathematics B	2014	2.29	5.22	0.42	30 484	F
B100	Biology A Modules B1, B2, B3	2014	1.37	3.08	0.41	15 694	F
B101	Biology A Modules B4, B5, B6	2012	2.34	7.84	0.37	1105	F
S108	Science Modules B1, C1, P1	2012	1.56	3.66	0.35	10 765	H
M108	Mathematics B	2012	1.60	3.80	0.34	18 784	F
S100	Science A Modules B1, C1, P1	2013	1.24	3.29	0.34	6730	F
B103	Biology A: Unit 1 (B1, B2, B3)	2010	1.53	3.85	0.33	4042	F
M108	Mathematics B	2013	1.99	4.86	0.30	27 972	F
S101	Science A Modules B2, C2, P2	2013	1.26	2.88	0.30	10 627	H
C107	Chemistry B: Unit 2 (C4, C5, C6)	2012	0.99	2.40	0.30	1304	F
S100	Science A Modules B1, C1, P1	2013	0.77	1.79	0.30	7982	H
S100	Science A Modules B1, C1, P1	2012	1.64	3.93	0.28	6862	F
P108	Physics Modules P4, P5, P6	2015	0.63	1.35	0.28	421	F
S108	Science Modules B1, C1, P1	2014	0.92	2.10	0.28	30 137	H
S109	Science Modules B2, C2, P2	2013	1.31	3.37	0.27	25 097	F
S104	Science A: Unit 2 (B2, C2, P2)	2011	1.74	4.47	0.26	17 360	F

Note: All 29 of these components were identified as outliers using Tukey's upper fence. The assessment codes were created by the researcher and are not the real codes.

A short history of the Centre for Evaluation and Monitoring (CEM)

Chris Jellis (Cambridge CEM)

The Centre for Evaluation and Monitoring (CEM), formerly the Curriculum, Evaluation and Management Centre (CEM) was acquired from the University of Durham in 2019 by a joint venture between Cambridge University Press and Cambridge Assessment. Since then, it has established itself in a unique role within the wider Cambridge organisation due mainly to its groundbreaking computer adaptive assessments for use in schools. What follows is not intended to be an exhaustive account of all the assessments created in the last 40 years of CEM, but more a focus on some of the highs (and lows) of major interest during that time. The history of CEM is an interesting one, emphasising as it does the crucial importance of diligent research and rigorous statistical analysis to back up the claims any assessment provider makes.

Beginnings

In 1981, Colin McCabe at Newcastle University won a contract to evaluate the Technical and Vocational Education Initiative (TVEI) in the North East of England. McCabe, with colleagues, established the Curriculum Evaluation and Management Centre to carry out this evaluation. TVEI was a government sponsored initiative designed to increase the uptake of work-related skills and qualifications. It was overseen by the Manpower Services Commission (MSC) to run in tandem with the newly created Youth Training Scheme (YTS) and gave rise to changes such as the establishment of BBC microcomputers in schools, along with the move to rebrand traditional subjects such as Woodwork and Metalwork as Design and Technology and Home Economics as Food Technology.

Among the staff of the newly formed CEM Centre was Dr Carol Taylor Fitz-Gibbon, a researcher and economist who had spent some of her early career in the USA and had an interest in demonstrating value in a fair way. In 1982, she was approached by a school governor who had a very simple question. The governor wanted to know whether the Mathematics A Level results from their school were good given their intake.

Carol realised that without equivalent data from other schools, the question could not reasonably be answered. She further realised that although A Level results had a strong effect on choice of profession and future career progression, very little research had been carried out in this area. To a researcher with a keen mind, it seemed an important question that needed answers. It became a significant feature of her later work.

COMBSE

In 1983, Carol established a research project named COMBSE (Confidential, Measurement Based, Self Evaluation) (Fitz-Gibbon, 1985) to find an answer to this intriguing question. The COMBSE project ran from 1983 to 1987. The plan was to collect O and A Level scores from local schools and pool the data to establish the link between the two examinations. She correctly predicted that the average O Level grade was the best indicator of each A Level grade and she also knew that the use of O Level results to predict likely A Level results could be a concern, because the O Levels were themselves the product of the schools. She therefore sought a measure of general ability and tried a number of standard high-level ability tests. None worked well, but she was able to use the International Test of Developed Abilities (ITDA) which was being developed under the auspices of the International Association for Educational Assessment (Fitz-Gibbon, 1996, p. 61). That worked as a good predictor when augmented with a vocabulary test.

In order to provide a more comprehensive picture of A Level success, students were asked about the ways in which they were taught and also invited to complete a questionnaire with closed and open questions about their feelings and attitudes. This comprehensive monitoring system produced a model for much of the subsequent monitoring projects developed at Newcastle and Durham.

COMBSE started with 12 schools agreeing to share their data, and when it came to an end it was being used by 47 schools. It was clear that as more schools contributed data to the project, the better was the outcome for all those involved. COMBSE had confined itself to reporting on A Level Maths and English results only. Could a new system be designed that could provide schools with information on a much wider range of subjects? It was now time to bring those skills and experiences gained from the TVEI evaluation and COMBSE together to create a wider reaching research project.

Alis

In 1989, Carol took over as Director of the CEM Centre and established a new school evaluation system to replace COMBSE. This new system was called the A Level Information System (Alis). In the same year Peter Tymms, a former teacher, and later to become Director of CEM, became the first Research Associate to work on the project.

The team were keen to build on the success of COMBSE, but it was clear that the use of O Level results to predict likely A Level results could be a concern, particularly as O Levels were imminently to be replaced by the General Certificate of Secondary Education, the GCSE. A measure of general ability that worked as a good predictor was therefore required. Some well-regarded assessments of general ability were tried, but none provided the predictive power required by the project. To this end it was decided that CEM should create their own bespoke measure of student ability. This new assessment, called the Test of Developed Abilities (TDA), proved to have a much greater predictive power and became a standard part of the CEM testing model, not only for Alis, but for other CEM assessments that were to follow. The Alis system came to be used widely in secondary schools, providing as it did a measure of student ability and a prediction of future A Level results, which were vital for schools that were increasingly being measured by their outcomes.

Yellis

Following the success of the Alis system, considerations were made to create a similar system for younger students aged 14–16 that used predictions of likely GCSE results as the outcome rather than A Levels. This system, consisting of a new assessment providing a measure of general ability and a prediction of GCSE grades, was piloted in 1990 under the name Yellis (Year eleven information system). The pilot proved to be a success and the assessment was released to schools in 1992.

PIPS and ASPECTS

In the same year, an assessment for children in Year 6 of primary school was started. The new system was called PIPS (Performance Indicators in Primary Schools) and was soon modified to cover all year groups from Year 1 to Year 6. These were designed by the PIPS Director, Peter Tymms (Tymms 1999), who wrote the initial tests including the PIPS Baseline assessment for 4–5 year old children starting school in 1993. He also designed the feedback given to schools.

In 1994, Christine Merrell[1] was appointed, and her particular interest in Early Years education led to the development of an assessment for 3–4 year olds in nurseries, ASPECTS. PIPS Baseline was used by a quarter of primary schools in England in 1998, some schools having joined the project as part of a government initiative of national testing in the early years. It was replaced when the Early Years Foundation Stage (EYFS) was introduced in 2008, which involved a very different kind of approach to assessment (QCA, 2008).

The CEM approach to assessment

Carol's early work established some basic principles. Her goal was to use effective psychometric models that are good predictors of future achievement to create assessments that are dependable and fair. Another aim was to reduce the burden of assessment on teachers and students, which led rapidly to the adoption of computer adaptive testing. The main aim was to use the data from these assessments to provide teachers and school leaders with valid and reliable data upon which to make their decisions. Finally, there was the fundamental belief that teachers and leaders were in the best position to decide what to do with the data for their school.

[1] Christine died recently after a short illness.

A new home

The organisation was growing and starting to have an influence upon school performance so, after some disagreements with Newcastle University in 1996, Carol was offered a new post at Durham University and moved the CEM Centre with her to the city of Durham, initially to offices close to the School of Education and then to larger premises on the Durham University Science Campus. Along with this success came more money, allowing the establishment of new posts, and among those appointed at this time were Robert Coe and Kate Bailey, both to become future directors of CEM. Also at this time, a pilot for a new assessment for students aged 11–14 in secondary school was launched. This assessment, known as MidYIS (Middle Years Information System) provided a measure of student ability, plus a prediction to GCSE.

InCAS

In 2002, CEM launched InCAS (Interactive Computerised Assessment System), a groundbreaking new computer adaptive assessment which used a single piece of software to cover the age range from 5 to 11. Data from each of the PIPS assessments in Years 1 to 6 (ages 5 to 11) were analysed using the Rasch statistical method, enabling the team to establish a single scale in each of the key cognitive areas for the whole primary range. These scales were then used to build a single computer adaptive test. Students would start the assessment with items easy for their age and through adaptive testing their ability level would be established and recorded as an age equivalent score. The system provided a reliable and efficient way of measuring student abilities. As students took the assessment each year, a measure of longitudinal progress of their time in primary school was established. InCAS went on to be adopted for a number of years as a mandatory assessment for use in state primary schools in Northern Ireland.

BASE

In 2015, the UK government planned to mandate a baseline assessment in the reception classes of English state schools, to provide teachers with a measure of what pupils knew and could do when they started school. CEM had been running the PIPS baseline assessment successfully for many years, and grasped the opportunity to develop a new baseline assessment along the general lines of PIPS but updated to take into account the feedback received from teachers and schools over this time. CEM was now under the directorship of Robert Coe and CEM's bid was successful. The subsequent assessment, known as BASE, became one of the mandated reception assessments for the next two years. After this time, government policy changed under pressure from unions and other lobbyists and mandated reception assessment was dropped (in 2021 it was reintroduced in yet another form). The BASE assessment, however, continues and is used around the world.

iPIPS

Although not a CEM commercial product, the iPIPS system was developed by Peter Tymms to provide information for policy makers about what is happening in the first year at school. It involved translating the original PIPS baseline assessment into many different languages The iPIPS system has been used to great effect in Brazil, Lesotho, South Africa and Russia, and the findings from those studies form the subject of a book (Tymms et al., 2023).

Check Together

The first assessment produced after CEM joined Cambridge in 2019 was a modified version of the BASE assessment specifically designed for use in Cambridge schools in India. This version, featuring a uniquely Indian soundtrack, imagery, content, and reports was developed in collaboration with colleagues in Cambridge.

The Cambridge Wellbeing Check

Following the Covid-19 pandemic in 2020, and the detrimental effects caused to school pupils due to school closures, greater emphasis started to be placed on student wellbeing than had previously been the case. The Cambridge Wellbeing Check was developed from a survey developed by researchers Dr Ros McLellan, Maurice Galton, Susan Steward and Charlotte Page in the University of Cambridge's Faculty of Education (McLellan & Steward, 2015). The original survey was created as part of a study examining the role of creative initiatives in fostering wellbeing, which was funded by the international creative learning foundation Creativity, Culture and Education. CEM has since worked with Dr McLellan and her colleagues to refine the questionnaire. It is now administered as a digital check for students aged 7 and above, alongside materials teachers can use to support school wellbeing initiatives.

Preliminary work is now being carried out to further integrate wellbeing with other CEM assessments and provide greater insights.

Cambridge Early Years Check Together

Following the development of Check Together in India, Cambridge colleagues requested a version of the assessment to augment the newly developed Cambridge Early Years curriculum. A new soundtrack, graphics and content were developed with the view to provide an assessment appropriate for as wide an audience as possible, along with greater integration with the Cambridge Early Years curriculum. The assessment was launched in the autumn of 2023.

Controversy

CEM's story has been intertwined with the Department for Education (DfE) and their initiatives for a long time, providing both support and challenge. Although originally established to evaluate the Technical and Vocational Education Initiative (TVEI), that evaluation and subsequent report (Fitz-Gibbon et al, 1988) found worse outcomes for those students that had been involved in the TVEI project than those that had not. Considering that the TVEI project had a budget of £900 million, this was quite a blow and was not received well.

Similarly, in 1999 the Education Secretary David Blunkett, hit out at CEM researchers (TES, 1999) who challenged the government view that older primary school children should be set 30 minutes of homework each night. CEM's research involving a survey of 20 000 pupils found that those who were set homework just once a month achieved better test scores.

In 2001, Professors Tymms and Fitz-Gibbon (2001) challenged the validity and accuracy of government figures regarding the increase in standards of Key Stage 2 results. Their work examined exam results over the previous 25 years and found some rise in standards, but not to the extent claimed by the government.

Again in 2004, Professor Tymms published an article in the *British Educational Research Journal* (Tymms, 2004) questioning the government's claims that literacy standards among 11-year-olds had risen dramatically between 1995 and 2000. This enraged the then Education Secretary, Ruth Kelly (Mansell, 2005), but Tymms' central argument was backed by the Statistics Commission, a non-departmental public body set up to oversee the work of the Office for National Statistics which refused to change its view, even in the light of heavy government pressure. The Statistics Commission's report (Statistics Commission, 2005) included a letter from Tim Oates, then head of research and statistics at the Qualifications and Curriculum Authority (QCA) which also supported Tymms' position.

Carol Fitz-Gibbon and Peter Tymms also came under pressure from statisticians to use multilevel models when analysing school data. In fact, Carol had considered using multilevel methods early in the development of the Alis assessment and wrote a paper discussing the use of such models (Fitz-Gibbon, 1991). Although acknowledging the strengths of the method, she ultimately rejected it for use in the Alis system as she felt that using a simpler system would be easier to explain to school personnel. Nevertheless, Carol Fitz-Gibbon and Peter Tymms were invited to explain their approach in a meeting at the Department for Education with Harvey Goldstein (a member of the Royal Statistical Society and a leading proponent of multilevel modelling), and Nick Tate (chief curriculum and qualifications adviser to the Secretary of State for Education). They were able to successfully argue their case.

> "Harvey said 'you've got to use multilevel models' and in fact we said 'no, no, no. If you look at the results in multilevel models, they are exactly the same as the ones you get out of classical tests' and we had a meeting, a showdown with Harvey at the DfE under Nick Tate and we won the argument against Harvey. I don't think we were ever forgiven for that." (Peter Tymms, personal communication)

Carol and Peter's work with Luke Vincent on the comparative difficulty of A Level subjects (Fitz-Gibbon and Vincent, 1994; Tymms and Vincent, 1995) resulted in further criticism from Harvey Goldstein and Michael Cresswell (Goldstein and Cresswell, 1996), this time focusing on their use of the subject pairs analysis approach and the use of Alis data in the analysis. The controversy continued for some time and was addressed again in 2008 (Coe et al., 2008) by a team led by Robert Coe, who went on to become the Director of CEM in 2010.

New ideas

Carol Fitz-Gibbon had previously worked in the USA and brought some of the prevailing ideas about education measurement with her when she returned to the UK. One of these ideas was the concept of value added. Following the success of Alis, she won a contract to set up a value-added system in Scotland for Highers using Standard Grade results as the baseline, which lasted for many years (Fitz-Gibbon, 1992).

This piqued the interest of the Westminster government. In 1995 it commissioned a contract to research a new model for measurement of school outcomes. CEM won the contract, and in 1997 the Value-Added National Project report was published (Fitz-Gibbon, 1997). The report recommended a method of determining value added and a variation on the general approach was then adopted by the government.

Carol was also a great advocate of the Randomised Control Trial (RCT) (where subjects are randomly assigned to one of two groups, experimental and control) and was influential in the creation of the Campbell Collaboration project in the USA. At the time it was extremely unusual to use RCTs in educational research but subsequently they were used to great effect by CEM staff in peer learning projects in Scotland (Tymms et al., 2011). It was the first randomised control trial for peer tutoring that went across a whole local education authority, and it is believed to have been the largest randomised control trial in education at the time. Now RCTs are widely used in education.

For many years in the UK, analysis of test results from examinations and other assessments used a model called Classical Test Theory (CTT). Carol realised that a newer model, called Item Response Theory (IRT) was being used extensively in other countries, particularly the USA and Australia. She advocated its use in the UK too but fell foul of some of the leading statisticians in the UK, who felt that the model was not appropriate (see for example Goldstein, 1979; Panayides et al., 2010). Undaunted, Carol continued and the IRT model is now used extensively in CEM assessments. To establish greater interest in IRT measurement, Peter Tymms held a meeting at Durham University of likeminded people who were working with the Rasch model, including Tom Bramley from Cambridge Assessment. This established the UK Rasch User Group, which has met regularly for many years, and of which Cambridge is a very active member.

Outreach

From its earliest times, CEM has had an effect, not only on education in the UK, but also around the world. In 1998, CEM established a relationship with the University of Western Australia and established a CEM outreach centre there with Helen Wildy as director. A year later CEM established another outreach centre in New

Zealand, followed by one in Hong Kong in 2001. These centres were able to foster regional interest in CEM assessments and research and reach a much greater audience than could be achieved from the UK alone.

Research

As CEM expanded, its research section grew accordingly. The section rapidly gained attention as a centre for excellence and won many contracts from organisations such as the Sutton Trust and the Education Endowment Foundation, contributing significantly to Durham University's research excellence framework (REF) submission. Many studies, such as the peer learning study in Fife, Scotland (Tymms et al., 2011) and various explorations into the nature of ADHD (Attention Deficit Hyperactivity Disorder) manifestation in the classroom (Sayal et al., 2020), have also used CEM assessments as pre- and post-measures of ability when investigating potential educational interventions.

Present day

Currently, all CEM assessments are delivered digitally, and work has been ongoing to explore how the capabilities in CEM can be brought to bear on enhancing the Cambridge offer to schools in the UK and overseas. Kate Bailey, who started in CEM in 1996, is now the Managing Director, replacing Elizabeth Cater who headed CEM after the integration with Cambridge. Current and previous CEM directors have recently published a book, *The First Year at School: An International Perspective* (Tymms et al., 2023), which details work on the iPIPS project and its effect around the world. The book is dedicated to Christine Merrell who created the PIPS and ASPECTS baseline assessments with Peter Tymms and created the original design for the BASE assessment.

CEM's focus for the future will be on strengthening the baseline assessments that CEM is known for and ensuring that they can support all Cambridge schools in improving the outcomes of learners all round the world. There is more to do in exploring how the unique capabilities in CEM can be used to enhance the Cambridge portfolio and reach even more learners in future.

References

Coe, R., Searle, J., Barmby, P., Jones, K., & Higgins, S. (2008). *Relative difficulty of examinations in different subjects*. CEM centre.

Fitz-Gibbon, C. T. (1985). A-level Results in Comprehensive Schools: the COMBSE project, Year 1. *Oxford Review of Education, 11*(1), 43–58.

Fitz-Gibbon, C. T. (1991). Multilevel modelling in an indicator system. In *Schools, classrooms, and pupils* (pp. 67–83). Academic Press.

Fitz-Gibbon, C. T. (1992). *Performance indicators and examination results.* Scottish Office Education Department, Research and Intelligence Unit.

Taylor Fitz-Gibbon, C. (1996). Monitoring Education. *Monitoring Education: Indicators, Quality and Effectiveness*. Bloomsbury.

Fitz-Gibbon, C. T. (1997). *Feasibility studies for a national system of Value Added indicators*. SCAA.

Fitz-Gibbon, C. T., & Vincent, L. S. (1994). *Candidates' performance in science and mathematics at A-level*. School Curriculum and Assessment Authority.

Fitz-Gibbon, C. T., Hazelwood, R. D., Tymms, P. B., & McCabe, J. J. C. (1988). Performance indicators and the TVEI pilot. *Evaluation & Research in Education, 2*(2), 49–60.

Goldstein, H. (1979). Consequences of using the Rasch model for educational assessment. *British Educational Research Journal, 5*(2), 211–220.

Goldstein, H., & Cresswell, M. (1996). The comparability of different subjects in public examinations: A theoretical and practical critique. *Oxford Review of Education, 22*(4), 435–442.

Mansell, W. (2005, August 26). *Why this man scares Ruth Kelly*. TES.

McLellan, R., & Steward, S. (2015). Measuring children and young people's wellbeing in the school context. *Cambridge Journal of Education, 45*(3), 307–332.

Panayides, P., Robinson, C., & Tymms, P. (2010). The assessment revolution that has passed England by: Rasch measurement. *British Educational Research Journal, 36*(4), 611–626.

QCA. (2008). *Early years foundation stage Profile handbook*. Qualifications and Curriculum Authority.

Sayal, K., Merrell, C., Tymms, P., & Kasim, A. (2020). Academic outcomes following a school-based RCT for ADHD: 6-year follow-up. *Journal of Attention Disorders, 24*(1), 66–72.

Statistics Commission. (2005). *Measuring standards in English primary schools*. Statistics Commission Report, 10.

TES. (1999, July 23). Too heavy with elite (p. 12).

Tymms, P. (1999). *Baseline assessment and monitoring in primary schools*. David Fulton Publishers.

Tymms, P. (2004). Are standards rising in English primary schools? *British Educational Research Journal, 30*(4), 477–494.

Tymms, P., & Fitz-Gibbon, C. (2001). Standards, achievement and educational performance: a cause for celebration? In R. Phillips, & J. Furlong (Eds.), *Education, reform and the state: Twenty-five years of politics, policy and practice*. Routledge.

Tymms, P. B., & Vincent, L. (1995). *Comparing Examination Boards and Syllabuses at A-level: Students' Grades, Attitudes and Perceptions of Classroom Processes: Technical Report*. Northern Ireland Council for the Curriculum, Examinations and Assessment.

Tymms, P., Bartholo, T. Howie, S. J., Kardanova, E., Koslinski, M. C., Merrell, C., & Wildy, H. (Eds.). (2023). The first year at school: An international perspective, *International Perspectives on Early Childhood and Development, 39*, Springer.

Tymms, P., Merrell, C., Thurston, A., Andor, J., Topping, K., & Miller, D. (2011). Improving attainment across a whole district: School reform through peer tutoring in a randomized controlled trial. *School Effectiveness and School Improvement, 22*(3), 265–289.

Research News

Lisa Bowett (Research Division)

The following reports and articles have been published since *Research Matters*, Issue 36:

Journal articles and other publications

Chambers, L., Vitello, S., & Vidal Rodeiro, C. (2024). *Moderation of non-exam assessments: a novel approach using comparative judgement*. Assessment in Education: Principles, Policy & Practice.

Constantinou, F., & Carroll, M. (2023). Teacher-student interactions in emergency remote teaching contexts: Navigating uncharted waters? *Learning, Culture and Social Interaction, 43*.

Constantinou, F. (2024). *'If you have a question that doesn't work, then it's clearly going to upset candidates': what gives rise to errors in examination papers?* Oxford Review of Education.

Majewska, D., Horsman, R., & Angove, J. (2024). Mapping HOTmaths Lessons to the Common Core State Standards for Mathematics. *Educational Designer, 4*(16).

Yu, J., Kreijkes, P., & Salmela-Aro, K. (2022). Students' growth mindset: Relation to teacher beliefs, teaching practices, and school climate. *Learning and Instruction, 80.*

Yu, J., Kreijkes, P., & Salmela-Aro, K. (2023). Interconnected trajectories of achievement goals, academic achievement, and well-being: Insights from an expanded goal framework. *Learning and Individual Differences, 108*.

Research and statistics reports on our website

Abu Sitta, F., Maddox, B., Casebourne, I., Hughes, S., Kuvalja, M., Hannam, J., & Oates, T. (2023). *The Futures of Assessment: Navigating Uncertainties through the Lenses of Anticipatory Thinking.*

Carroll, M. (2023). *Sex gaps in education in England*.

Williamson, J. (2023). *Cognitive Diagnostic Models and how they can be useful*.

Williamson, J., & Vidal Rodeiro, C.L. (2024). *Progression from GCSE to A level, 2020–2022.*

Conference presentations

The annual conference of the Association for Educational Assessment – Europe (AEA-Europe) took place in Malta, 1 to 4 November 2023, https://2023.aea-europe.net. Our researchers presented a total of seven papers:

Constantinou, F. *Can exam papers always be error free? An exploratory investigation into the conditions that can give rise to errors in assessment instruments.*

Ireland, J., & de Groot, E. *Multiple marking using the Levels-only method for A level English Literature*.

Kuvalja, M. *Evaluation of the Cambridge International Digital Mock Exams Service*.

Rushton, N., & Lestari, S. *COVID-19 related changes to upper secondary assessments in six countries: Adaptations and reactions*.

Vidal Rodeiro, C. L., & Williamson, J. *Evaluating the impact of curriculum and assessment reform in secondary education on progression to mathematics post-16.*

Walland, E., & Leech, T. *How are GCSE grades used in post-16 admissions decisions in England?*

Williamson, J., & Vidal Rodeiro, C. L. *Performance in secondary mathematics topics pre- and post-reform*.

Blogs and insights

Johnson, M. (2023, November 17). *'Doing time': Issues for qualitative research when dealing with data related to time*. British Educational Research Association Blog.

Johnson, M., & Cambridge International Education. (2023). Getting started with... oracy. *Cambridge International Education Teaching and Learning Resources.*

Kuvalja, M. (2023, November 29). *The use of ChatGPT for content creation: A student perspective*.

Moran, R. (2023). Teachers' views on access arrangements. (Based on Vidal Rodeiro & Macinska, 2003, published in *Research Matters, 35*, 41–59)

Vitello, S., Majewska, D., & Walland, E. (2024, February 01) *The power of research - why is it important for assessment?*

Sharing our research

We aim to make our research as widely available as possible. Listed below are links to the places where you can find our research online:

Journal papers and book chapters: https://www.cambridgeassessment.org.uk/our-research/all-published-resources/journal-papers-and-book-chapters/

Research Matters (in full and as PDFs of individual articles): https://www.cambridgeassessment.org.uk/our-research/all-published-resources/research-matters/

Conference papers: https://www.cambridgeassessment.org.uk/our-research/all-published-resources/conference-papers/

Research reports: https://www.cambridgeassessment.org.uk/our-research/all-published-resources/research-reports/

Data Bytes: https://www.cambridgeassessment.org.uk/our-research/data-bytes/

Statistics reports: https://www.cambridgeassessment.org.uk/our-research/all-published-resources/statistical-reports/

Blogs: https://www.cambridgeassessment.org.uk/blogs/

Insights (a platform for sharing our views and research on the big education topics that impact assessment around the globe): https://www.cambridgeassessment.org.uk/insights/

Our YouTube channel: https://www.youtube.com/channel/UCNnkOpi7n4Amd_2afMUoKGw contains Research Bytes (short presentations and commentary based on recent conference presentations), our online live debates #CamEdLive, and podcasts.

You can also learn more about our recent activities from Facebook, Instagram, LinkedIn and X (formerly Twitter).

Contents / Issue 37 / Spring 2024

Cambridge University Press & Assessment
Shaftesbury Road
Cambridge
CB2 8EA
United Kingdom

ResearchDivision@cambridge.org
www.cambridge.org