CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Research Matters

Issue 38 / Autumn 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: Constantinou, F. (2024). Troubleshooting in emergency education settings: What types of strategies did schools employ during the COVID-19 pandemic and what can they tell us about schools' adaptability, values and crisis-readiness? *Research Matters: A Cambridge University Press & Assessment publication, 38*, 6–27.
https://doi.org/10.17863/CAM.111626

Credits

Reviewers: Jo Ireland, Carmen Vidal Rodeiro, Matthew Carroll, Martin Johnson and Nicky Rushton

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 38, https://doi.org/10.17863/CAM.111625

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

Contents

Foreword

Tim Oates, CBE

With 2024 as the third year in which exams were taken in lower and upper secondary, there is a pervasive sense of "getting back to pre-pandemic arrangements". But with COVID-19 impact still moving up through the system, a new UK government, and reflection on how the nation managed interrupted education, questions continue to be asked about international comparisons, the amount of formal assessment, and the resilience of systems. Whether the questions are old and familiar or new and challenging, we all benefit from evidence-based answers to them. The articles in this edition of *Research Matters* are central to further development of assessment and learning. Within existing examination approaches can we reduce the assessment duration whilst preserving the integrity of measurement? How can post-16 qualifications support those who, at the end of lower secondary, fall below national expectations in national exams in maths? What impacts are arising from the general societal switch from handwriting to typed script? And with concerns about individual welfare and wellbeing amplified during the pandemic, can we learn from the experience of emergency? Proposals for change are coming thick and fast – some stimulated by the impact of COVID-19 and some nascent in the system from before the pandemic. I can't think of a more far-reaching natural experiment than pulling almost all children out of school on a punctuated and erratic schedule. The assessment research community continues to run hot trying to run national exams, understand what has happened over the past five years, and service calls for changed arrangements. Let's all work to keep both the profession and policy makers serviced with high-quality evidence which relates to the questions that matter most.

Editorial

Victoria Crisp

Welcome to the autumn issue of *Research Matters*. Disruption to school life due to emergency situations can have a substantial effect on students in terms of their learning and wellbeing. There is a need for schools to develop strategies for addressing the challenges of such situations. Our first article explores this issue in the context of the COVID-19 pandemic. Through a detailed analysis of interviews with teachers, Filio Constantinou identifies a set of macro- and micro-strategies used by schools and discusses the broader implications for emergency readiness.

In our second article, Tom Benton addresses a decision that has to be made about any high stakes test for any qualification type and subject: how long the test should be. Tom looks at the recommendations for minimum levels of reliability from the literature and uses these and psychometric formulae for the relationship between reliability and test length to calculate possible recommendations for test durations. To provide perspective, these recommendations are then compared to the test durations used in practice in a number of assessment contexts in different jurisdictions.

The importance of maths to everyday life and career opportunities continues to motivate moves to encourage greater uptake of maths among 16- to 18-year-olds. In our third article, Tim Gill looks at one example of this: Core Maths qualifications. These qualifications, which were first assessed in 2016, are targeted at learners who achieve at least a grade 4 in GCSE Maths but who do not go on to study A Level Maths. Tim's analysis explores the background characteristics of those taking Core Maths, the other qualifications and subjects these learners also study, and whether taking Core Maths is associated with better results in other qualifications with a quantitative element.

Moves towards increased use of digital examinations in general qualifications raise many interesting issues in relation to comparability. One important theme is how typing or word-processing responses, instead of handwriting them, may affect comparability. In our fourth article, Santi Lestari reviews the existing literature on the comparability of typed and handwritten long-answer responses, in terms of scores, text characteristics, marking, and composing processes.

In our final article, Lucy Chambers, Emma Walland and Jo Ireland describe a study in which they explored the use of the Comparative Judgement method for the assessment of music compositions and performances based on audio recordings. This is a novel use of Comparative Judgement as the method has primarily been applied to learners' written work. Lucy and colleagues use questionnaire data to investigate how judges felt about comparing music audio recordings and the factors influencing decision-making in this context.

Troubleshooting in emergency education settings: What types of strategies did schools employ during the COVID-19 pandemic and what can they tell us about schools' adaptability, values and crisis-readiness?

Filio Constantinou (Research Division)

Introduction

Crises, such as wars, epidemics, wildfires, earthquakes, and hurricanes, can disrupt education in major ways. When such crises occur, schools need to take immediate action[1] to prevent or mitigate any negative effects on student learning. With developments such as climate change and heightening geopolitical tension across the world increasing the frequency of crises globally (see e.g., Acevedo & Novta, 2017; Haileamlak, 2022; Senthilingam, 2017), it is crucial that schools become crisis-ready. Crisis-readiness lies partly in the ability of schools to deliver "emergency education" promptly and effectively. While originally linked to contexts impacted by armed conflict and its consequences (e.g., population displacement) (Kagawa, 2005), emergency education is currently understood as "education in situations where children lack access to their national education systems, due to man-made crises or natural disasters" (Sinclair, 2001, p. 4). Overall, it is seen as an emergency solution aimed at enabling teaching and learning to continue during a disruptive event such as a war, an earthquake, a flood or even an epidemic.

However, for emergency education to function as an effective solution, it needs to be informed by both an understanding of the educational challenges created by the disruptive event, or crisis, and an awareness of the types of strategies which could be employed to address these challenges. To support the delivery of

1 While in this study schools take action to address a crisis situation (i.e., schools as part of the solution), it should be acknowledged that sometimes schools can also play a role in the emergence of a crisis (i.e., schools as part of the problem). For example, for the dialectical relationship between formal schooling and armed conflict and the role of schooling in exacerbating inter-group hostility, see Kagawa (2005).

emergency education, this study focused on the latter dimension, that pertaining to possible courses of action during a crisis. Specifically, it sought to explore the kinds of resources that schools may mobilise and the types of measures that they can put in place to support their students when emergencies arise.

The crisis that provided the setting for this study was the COVID-19 pandemic. The COVID-19 pandemic, declared in 2020 (World Health Organization, 2020), caused an unprecedented disruption to education worldwide. According to the OECD (2022), in 2020, "1.5 billion students in 188 countries and economies were locked out of their schools" (p. 23). To curb the spread of the virus, many governments across the world imposed lockdowns, which resulted in schools closing for certain periods of time. Apart from learning loss (see e.g., Carroll & Constantinou, 2022; Di Pietro, 2023; Newton, 2021), the school closures also caused mental health and wellbeing problems among many students (see e.g., Deng et al., 2023; Panchal et al., 2023), exacerbating the overall negative educational impact of the disruption. The adverse effects of the pandemic continued even after schools reopened. For example, student and teacher absences increased as a result of quarantine rules, with teachers and students physically present at school having to operate in a highly unnatural pedagogical setting created by the social-distancing measures in place (see e.g., Howard et al., 2021; Sharp & Skipp, 2022).

This wide range of challenges triggered a number of different responses from schools, all intended to minimise the impact of the disruption on students or support students' recovery from the consequences of the crisis. To date, there have been various research attempts to capture these responses, or strategies, albeit probably not as many as those focusing on capturing the challenges. Common strategies reported in the literature include: providing students with academic support in core subjects either on a one-to-one basis or in small groups; adapting the curriculum; restructuring the school day; and offering pastoral support to students experiencing mental health difficulties (see e.g., Achtaridou et al., 2022; Bond et al., 2021; Crossfield et al., 2023; Johnson, 2022; OECD, 2022). This study sought to build on this research. Its aim was twofold: (a) to identify and document some of the strategies employed by schools, and (b) to illuminate their nature in order to gain insight into schools' adaptability and readiness to cope with a public health crisis such as that caused by the COVID-19 pandemic.

Methodology

To investigate the strategies used by schools during the COVID-19 crisis, this qualitative study drew upon in-depth interviews with 13 teachers based in different parts of Europe. The interviews were conducted as part of a larger mixed-methods project aimed at understanding the educational impact of the pandemic. The project involved a questionnaire completed by teachers based in different parts of the world, and follow-up interviews with 13 of them (for more information about the project, see Carroll & Constantinou, 2022, 2023; Constantinou, 2023; Constantinou & Carroll, 2023). The teachers participating in the project taught in schools that worked with the Cambridge Centre for Evaluation and Monitoring.

The interview participants were drawn from the pool of questionnaire respondents. They were purposefully selected to represent a heterogeneous group to allow a range of perspectives and experiences to be captured. As shown in Table 1, the interviewees worked in different education sectors (early years, primary, secondary), were based in different European countries, taught different subjects, and held different roles within their school. Their teaching experience ranged from six to 35 years. It is worth noting that, while both state and private schools were represented in the interview sample, the majority of the interviewees worked in the private sector.

The interviews, which aimed to provide more in-depth information about how schools from around the world experienced the COVID-19 disruption, were carried out online in June and July 2021. They were conducted in English. The choice of language did not create any communication issues, as the teachers who were not native speakers of English worked in (partly or fully) English-medium schools and were therefore fluent in English. In the interviews, which were semi-structured, the participants were invited to describe the challenges they faced during the pandemic and any strategies that they, or their schools, employed to address the implications of the crisis. Each interview lasted approximately 90 minutes. In line with the ethical guidelines for conducting educational research, written informed consent was obtained from all interviewees (see BERA, 2018).

The interview transcripts were subjected to thematic analysis (Braun & Clarke, 2021) using MAXQDA (VERBI Software, 2021). The analysis consisted of two stages. The first stage was more descriptive and focused on identifying the different micro-level strategies used by schools. The second stage, which was more interpretive, aimed to make sense of these micro-level strategies. This latter analysis stage, which was predominantly data-driven, led to the identification of a number of overarching mechanisms, or macro-level strategies, employed by schools to address the pandemic challenges. Both the micro-level and the macro-level strategies are explained and exemplified below.

Table 1: Interview participants (N=13)

Characteristics		N
School location	UK	7
	Cyprus	1
	Italy	2
	Romania	1
	Spain	1
	Switzerland	1
Education sector	Early years	1
	Primary	2
	Secondary	10
School type	State-funded	3
	Independent	10
Gender	Female	8
	Male	5
Position in the school	Teacher with a leadership role (e.g., head of department)	8
	Teacher without a leadership role	5
Subject area*	Creative subjects (e.g., art, design and technology, music)	2
	Humanities and social sciences (e.g., English language, literature, history)	5
	Science and mathematics	3

* This category concerns only the secondary teachers (the early years and primary teachers taught all subjects).

Findings

Overall, eight macro-level strategies were identified through the analysis. As shown in Table 2, these were organised into three groups based on three criteria: (a) the type of challenge targeted (the "what"); (b) the intended function (the "why"); and (c) the type of problem-solving approach employed (the "how").

Based on the first criterion, that is, the type of challenge targeted by the micro-level strategy, three kinds of macro-level strategies were identified:

- *Safety strategies:* These encapsulate safety measures put in place to reduce the risk of infection by the virus and enable school activities to be carried out safely.
- *Learning strategies:* These involve steps taken to support students' academic development which was disrupted by the pandemic.
- *Wellbeing strategies:* These entail actions intended at supporting students' mental health and overall wellbeing which also seemed to be affected by the crisis.

Based on the second criterion, that is, the intended function of the micro-level strategy, two varieties of macro-level strategies were detected:

- *Defence strategies:* These were aimed at providing protection against the crisis, by either averting or weakening the threat posed by it.
- *Recovery strategies:* These were employed to mend, or reverse, any harm caused by the crisis.

Finally, based on the problem-solving approach reflected in the micro-level strategy, three clusters of macro-level strategies emerged:

- *Suspension of existing structures:* Pausing activities which could no longer be carried out safely.
- *Exploitation of existing structures:* Using usual practices involving tools and resources already available in the school to combat the crisis or enable students to recover from it (e.g., incorporating more collaborative classroom tasks than usual into teaching).
- *Development of new structures:* Devising new, often creative solutions to address the challenges caused by the crisis. New structures took a variety of forms – some involved using existing tools and resources in new ways (e.g., converting a changing room into a temporary classroom), while others involved new tools or resources (e.g., face masks).

Each micro-level strategy received, overall, three attributes (through being linked, or assigned, to three macro-level strategies), one for each criterion, namely, the "what", the "why", and the "how", respectively. This enabled each micro-level strategy to be described and profiled. The micro-level strategies can be found in the following sections. For ease of presentation, they are organised into three sections based on the first criterion, that is, the type of challenge targeted (i.e., the "what"). Each section concludes with a summary table which captures the profile of each micro-level strategy.

Table 2: The eight macro-level strategies

Macro-level strategies	
The "what" (=type of challenge targeted by the micro-level strategy)	Safety strategies
	Learning strategies
	Wellbeing strategies
The "why" (=intended function of the micro-level strategy)	Defence strategies
	Recovery strategies
The "how" (=type of problem-solving approach reflected in the micro-level strategy)	Suspension of existing structures
	Exploitation of existing structures
	Development of new structures

Safety strategies

As explained below, the safety strategies reported in the interviews took the form of a series of defence measures employed by schools when in-person instruction resumed.

Defence strategies

To protect themselves against the virus and reduce the risk of infection, schools either paused teaching and learning activities ("suspension of existing structures") or invented new solutions to allow such activities to continue ("development of new structures"). Interestingly, no evidence of attempts to draw upon existing mechanisms ("exploitation of existing structures") was detected. This is indicative of the absence of such mechanisms which is, in turn, suggestive of a lack of preparedness on the part of schools to cope with the safety challenges of such a crisis.

Examples of activities which were suspended due to being deemed unsafe after schools reopened, included singing, group art projects, and school assemblies:

> "Whereas now, [music] lessons are static, and they [students] come in and they sit at a table. And they can't sing. I can't sing to demonstrate anything." [UK]

> "Because of COVID I have not planned [art and design] group work. However, I did do a group work halfway through the year with making bridges, and they had to work in pairs to do that. It was such a pain, I had to disinfect everything all the time. Now, everybody's working on their own thing. I decided not to do another group work, just to make it easier." [Romania]

> "Normally, schools have assemblies for Year groups when you talk about different topics and the news, anything that is going on – all of these have been cancelled." [Spain]

To enable as much teaching and learning to continue as safely as possible, schools attempted to implement social distancing where feasible. Social distancing, a public health practice intended to reduce the rate of virus transmission through

minimising close physical contact between individuals, was a governmental requirement with which schools had to comply. The social-distancing measures employed by schools took a variety of forms, ranging from implementing relatively small changes to student and teacher practices, to undertaking more radical interventions. The former included introducing new routines such as wearing a mask, disinfecting hands and surfaces, moving activities outdoors where possible, and walking in designated areas indicated by floor markings or other signs:

> "Wearing masks all the time. Because of the space, it was compulsory." [Spain]

> "You say 'OK, we're about to get out the instruments, so here's the hand sanitiser again, here's some wipes. We're all going to clean the beaters'." [UK]

> "My school is close to a big park, so in the sunny days we delivered our lessons in this park." [Italy]

> "So, in terms of distancing in the classroom, the teacher is to be two metres from the pupils. So, in the classroom we have these physical strips of tape on the ground to show the separation." [UK]

Another defence strategy employed concerned the classroom seating arrangements which some schools amended to facilitate social distancing. As reported in the interviews, students sat in rows rather than in groups, often on their own, with plastic barriers sometimes separating them from their classmates:

> "I suppose one thing is we've had the children in rows in the school, which isn't normal for primary schools. Normally, it's sitting them around in groups." [UK]

> "And when we came back, the classes were small enough that we could put a child and then a gap, and then a child and then a gap – so we had enough space to do that." [Switzerland]

> "Now we have separate desks for each student. And each desk is with sort of a cover made of plexiglass." [Italy]

In some cases, students also had to sit in an alphabetical order, a measure normally taken to help simplify contact-tracing efforts:

> "A decision that has been made in our schools is that all pupils sit in alphabetical order for the entire year, which again isn't ideal because usually you would change the seating plan at least three or four times a year." [UK]

An additional defence strategy involved operating a "bubble" system. This entailed organising students into "bubbles", that is, into smaller clusters (e.g., based on their Year group or Key Stage group). Students in each bubble remained together for most, or all, activities throughout the day (e.g., lessons, breaks), avoiding interaction with students outside of their bubble. This was intended to

reduce the risk of virus transmission and allow teaching to be delivered in a safer way:

> "So, the classes, the Year groups, are in bubbles. So, reception is in a bubble. Year 1 is in another bubble. The bubbles are not meant to mix. So, we do not gather for anything, like an assembly, and the children do not play in the same spaces, they have their own space." [UK]

> "We had Key Stage bubbles more or less [...]. So, the bubbles ate lunch at different times and had break in different places." [Switzerland]

To enable the bubble system to operate smoothly and fulfil its mission, schools created a zone system. In some cases, this involved building new cafeterias, each catering for a different bubble:

> "We have a café on site, but they actually built two other little, mini cafés, so that there was one for Key Stage 3, one for Key Stage 4, one for Key Stage 5." [Cyprus]

In cases where class size exceeded the limit set by the government, classes were split into two. To cater for the teaching needs of the additional classes, schools devised new, creative solutions. As explained below, one school in Switzerland resorted to a form of on-site synchronous hybrid teaching: teachers taught one half of the class in person, with the other half of the class attending the lesson online from a different room in the school:

> "So, my Year 12 class is my biggest class. It has 19 students, and I had ten students in the room, and nine students in another room that were online. So, we were doing hybrid teaching within the school. [...]. So, they alternated, so Monday I saw one half, on Tuesday I saw the other half. And they were all doing that – if your class was too big, then you had to separate." [Switzerland]

Splitting classes into sub-groups created a demand for more classrooms. Some schools addressed this need through converting other school areas, such as corridors, changing rooms, assembly halls and labs, into classrooms or study areas:

> "We have been teaching everywhere. We have been teaching in the large corridor, we have been teaching in a corner there, we have created classrooms from thin air. We have been teaching in the changing rooms, yeah." [Spain]

> "We converted the main hall into a study area for the sixth formers because there was a limit [of 20 people]. We couldn't use the main hall because it would have been difficult to maintain social distance, so we had no assemblies or anything. So, they converted the main hall into a study area for the sixth formers, with socially-distanced desks." [Cyprus]

> "Sometimes we had to teach in labs." [Cyprus]

Finally, infected students and teachers and their close school contacts were required to self-isolate at home for a period of time to prevent the spread of the virus:

> "We have, on occasion, lost parts of bubbles. So, we've never had to send an entire Year group home, but we have had big chunks of Year groups that have had to go home." [UK]

A summary of all the safety strategies identified in the study can be found in Table 3.

Table 3: The profile of safety strategies

Micro-level strategies	The "why"	The "how"
Suspended singing during music lessons.	Defence	Suspension of existing structures
Suspended group art projects.		
Suspended student assemblies.		
Introduced mask wearing.	Defence	Development of new structures
Introduced disinfection routines.		
Moved teaching outdoors (where possible).		
Introduced floor demarcation to encourage social distancing.		
Amended seating arrangements to facilitate social distancing and contact tracing.		
Operated a "bubble" system to reduce virus transmission.		
Operated a zone system.		
Split classes into two and implemented on-site synchronous hybrid teaching.		
Converted various school areas (e.g., corridors, changing rooms, assembly halls) into classrooms to facilitate social distancing.		
Required infected students and teachers and their close school contacts to self-isolate at home for a period of time.		

Learning strategies

Another type of challenge confronted by schools during the pandemic was supporting student learning. As the analysis indicated, this support took two forms: (a) reducing the risk of learning loss caused by the disruption (defence strategies), and (b) helping students to catch up on any learning they may have missed due to school closures (recovery strategies).

Defence strategies

It is worth noting that all learning-related defence strategies reported in the interviews involved developing new structures. As there did not seem to be any school structures in place which could be exploited to reduce the negative impact on student learning, new and creative solutions had to be devised.

To minimise learning loss during lockdown, many schools around the world switched to remote teaching which, in many cases, took the form of live online lessons:

> "We were doing live lessons [online] all the way through those national lockdowns." [UK]

In cases where some students could not attend the live online lessons – either because they were unwell or because they were based in a different time zone – teachers video recorded the lessons to prevent them from falling behind:

> "When we were all remote, we had to record all of our lessons online, and so some pupils watched them on record at different times of day, depending on where [in the world] they were." [UK]

Online instruction was a novel experience for most teachers and involved a number of challenges which often undermined or complicated the process of teaching and learning. Key challenges included:

- Students were more susceptible to becoming passive and disengaged during online learning (Challenge 1).
- It was difficult for teachers to know whether, or to what extent, students were able to follow the online lesson and understand what was being taught, as they could not see them (Challenge 2).
- Some students did not have access to the necessary learning resources and materials during lockdown (Challenge 3).
- It was challenging for group work to be carried out effectively online (Challenge 4).

To support student learning as much as possible during online lessons, teachers developed various strategies. These are summarised in Tables 4 to 7. They are presented based on the type of challenge they sought to tackle and are all exemplified through relevant interview extracts.

Table 4: Strategies developed to address Challenge 1 (=Students were more susceptible to becoming passive and disengaged during online learning)

Strategies developed to address Challenge 1	**Interview extracts (quotations)**
Calling students at random during the lesson to encourage them to be present and attentive.	If the kids don't have their cameras on, you don't know whether they're even present. I would, halfway through a lesson, start calling on random students and I would tell them I was going to do that to check that they were present. So, you had to employ whatever means you could to make sure that they were attending. [UK]
Assigning tasks that involved a physical element (e.g., writing or drawing on paper) to prevent students from becoming passive.	I would then set an open-ended task [...], say, a writing activity, and I was a great fan of the "hold it up and show me what you've done at the end", so they've got to actually have physically done something. We try to get them off screens as much as possible so they were actually writing something and drawing something, rather than just typing or accessing passively the screen. That was one of the things that we set out from early on, because we could see a danger in that, the children just becoming totally passive. [UK]
Making the lesson as enjoyable as possible for the students.	But I tried to keep as much of those fun things, the things they like. They like a little two-minute or three-minute film, or they like the opportunity to not just listen to me. [UK]
Asking students differentiated questions in the chat and encouraging them to respond to, or challenge, another student's answer.	Often a strategy I used, again for the chat, maybe I would write down a set of questions, differentiated questions, so a different question for each pupil. So, depending on their level, I'd ask them a more challenging or less challenging question, and ask them to respond in the chat, and then I would give them some time to look at each other's responses, and then quite a nice tactic was sometimes to ask each of them to respond to or challenge one other person's response. [...]. So, by asking everyone to respond in the chat, it means everyone was involved. [UK]

Table 5: Strategies developed to address Challenge 2 (=It was difficult for teachers to know whether, or to what extent, students were able to follow the online lesson and understand what was being taught, as they could not see them)

Strategies developed to address Challenge 2	**Interview extracts (quotations)**
Reducing taught content.	We cut content when we were online. [...]. So, we took content out that we felt wasn't absolutely necessary to be taught at that particular point. [UK]
Simplifying the lesson and focusing only on key points.	If I'm fully online teaching, I'm keeping things a lot simpler. There's less room for complication. [UK] So, I tried to stick to the main points. [Italy] So, I felt that teaching became linear, you know, everything had to follow a straight line. It was hard to go off on a tangent, or if there was something that was particularly interesting that had been thrown up, it was hard to then address that, because you didn't know whether you were just talking to one student rather than having everybody on board. [UK]
Slowing down the pace of teaching.	And I think I probably am now more actively aware that sometimes it's more important to slow down and ensure that everyone is keeping up that greater quality, than just blindly running through the material and hoping that people catch up. So, I think I've simplified and slowed down and seen the value in maybe quality over quantity. [UK] I tried to be very, very slow in teaching because I get that for my students it was difficult to follow the entire lesson online. [Italy]
Communicating clearly and explicitly.	Because it's very easy to miss stuff online and it's very easy for the teacher not to see that people haven't grasped what's required, you had to be really specific in laying out what the requirements were. [...] but you had to be doubly sure when you were online that everyone knew what was coming and what was expected of them. So, explicit, clear instructions were absolutely essential. [Italy]
Taking detailed notes of ideas mentioned in the class to enable everyone to follow the lesson.	So, I found myself writing a lot of notes onto the class notebook to annotate ideas that were coming from class discussion or to type my own ideas. So, instead of perhaps my writing a couple of notes on a whiteboard, and then being able to check organically in the classroom whether pupils understood or not, I found myself writing much, much more, just to be absolutely sure that everybody was keeping up. That was also important because I did have two girls in that class who were learning through recordings. So again, I wrote everything down, everything useful to make sure that they were keeping up. [UK]

Table 6: Strategies developed to address Challenge 3 (=Some students did not have access to the necessary learning resources and materials during lockdown)

Strategies developed to address Challenge 3	Interview extracts (quotations)
Choosing to teach topics that did not require specialised material to allow all students to participate in the lesson.	It was more to do with the fact that most students don't have more than a pencil at home, right? I would say only half of them had colour pencils and watercolours. That's why I ended up doing perspective, because I thought most people will have a pencil and a ruler, you'd think, right? I said "If you don't have a ruler, use a book." [Laughs]. This is what I'm dealing with. All my teaching this year has been just pencil and paper really. Normally, I'd be painting and I'd be making sculptures, we'd be doing all sorts of things. But I just can't do it, I can't do it when half the class doesn't have the materials. [Romania]
Providing students with a digital version of any necessary learning resources, where possible.	Most of them had the books that they needed at home, but some of them didn't, and so I literally took pictures on my phone and would send them pictures of the music so that they could do it. [UK]

Table 7: Strategies developed to address Challenge 4 (=It was challenging for group work to be carried out effectively online)

Strategies developed to address Challenge 4	Interview extracts (quotations)
Increasing group size to account for the likelihood of some students leaving the breakout room because of technical issues.	And, also, issues with connectivity. If you have a group of three, and two people lose connection, then you've got one person on their own. So really, for the breakout rooms to work, I was having to put students into groups of at least four, whereas normally I would – for me, that would be quite a big group to use. I think when you've got groups of four, you often get one person, at least, who isn't really contributing much. [Cyprus]
Designing shorter and more focused group tasks.	I have to be very careful with group work. [...]. So, if I set group work, [...], it'll be a much smaller task. You might say "Right, I'm going to get you in groups. You're going to read this, and you've got three minutes to come up with answers to this, this and this". So, they have to be quite carefully planned and focused. [Cyprus]

After schools reopened, the learning challenges remained but they manifested themselves somewhat differently. For example, due to quarantine rules, infected students, or close contacts of infected students, had to self-isolate at home and therefore miss school. To mitigate the risk of learning loss for the affected individuals, many schools implemented synchronous hybrid teaching as a defence strategy, to allow self-isolating students to continue attending lessons:

> "In our Sixth Form, we had students who were out in both Years 12 and 13 who were quarantining at home [...], and who weren't coming into the class. So, we were teaching hybrid classes where some of the students were live in person and some of the students were remote." [UK]

Sometimes, the individuals self-isolating were the teachers. To enable student learning to continue, self-isolating teachers – where possible – delivered lessons remotely from home. In this case, a teaching assistant would be present in the classroom to support the instructional process:

> "But many times, they [teachers] were just with mild symptoms or quarantining, so they were able to deliver their lessons via Zoom. And individuals like previous students of the school or people training in universities, came to the school and looked after the pupils whilst the lesson was delivered via Zoom by the teacher." [Italy]

However, it was not only the quarantine rules that jeopardised student learning after schools reopened. In some countries, transport restrictions constituted an additional source of disruption, preventing a subset of students from physically going to school. To address this issue, some schools developed a system whereby students were divided into three groups, each taught through a different medium: one group received fully in-person teaching, another group received fully online teaching while based at home, with the third group participating in synchronous hybrid lessons. For fairness, these groups alternated to allow all students to receive the same amount of in-person instruction:

> "So, since the buses and the underground are allowed only 50 percent of their capacity, we're supposed to split our classes into smaller groups. [...]. So, I had the three ways – the hybrid, the totally online, and the classroom [in-person] activity, because the groups swap. So, maybe one week one class was entirely at school, the other one was entirely at home, and we had hybrid ones as well. And then we moved to another pattern to give all the students the same possibility of attending the same number of lessons at school." [Italy]

Recovery strategies

Unlike the defence strategies which involved schools moving beyond existing arrangements and devising new solutions to support student learning ("development of new structures"), the recovery strategies drew upon established school mechanisms and already available resources ("exploitation of existing structures"). Employed mainly after the first phase of the disruption (i.e., after the first lockdown), the recovery strategies were aimed at reversing some of the harm caused by school closures by addressing gaps in students' knowledge and skills.

In the first instance, schools sought to collect information about their students' learning needs. They did so via conventional routes such as conducting student assessment and contacting parents:

> "What we've done now that they're all back is we've tested them towards the end of the year in a more rigorous way, and I think that will inform us where they are more accurately." [Italy]

> "We have a parent survey. So, I think that'll be a useful way of getting information from parents about what they know that their children are doing." [UK]

One strategy employed by schools to support the development of skills which had overall declined was making curriculum and pedagogy adjustments. For instance, a reception teacher incorporated more collaborative activities in her in-person lessons to help her young students to recover the social skills that they had lost during lockdown, while a secondary English teacher decided to place more emphasis on reading skills through reintroducing guided reading into her teaching:

> "And then, the main thing has been the social side, in terms of reminding children how we listen to each other. So, there's been more effort to build that up to remind everybody that we're part of a group again. So, a few more little social activities have been integrated because that's where the gap was. [...]. And so, they worked together in a group, collaboratively, on an art project. So, I've been thinking more about collaborative play, expressive arts, role-play type activities – thinking more of activities which unite children into groups again." [UK]

> "I know in the English department, I'm putting a lot more focus on reading and literacy next year, so we're bringing back library lessons. I'm looking at doing guided reading." [Cyprus]

A major focus of schools after reopening was providing support to students of lower academic ability who seemed to have been more severely affected by the disruption. This support took a variety of forms, notably a greater differentiation of teaching and provision of catch-up classes during or after school time:

> "So, for example, the lower ability [reception] children have had more practical activities within their programme. So, they've not just been writing the numbers to 10, they've been counting the teddy bears and the beads. They've been making patterns, lines, shapes. So, the lower ability have had more rich activities put into their programme to help them to deepen their understanding and improve their skills." [UK]

> "We've got intervention groups. We'll take groups who we perceive as being weak in a certain area, maybe spelling, punctuation, maybe grammar, maybe maths, whatever it might be, and we do catch-up groups." [UK]

A summary of the learning strategies can be found in Table 8.

Table 8: The profile of learning strategies

Micro-level strategies	The "why"	The "how"
Switched to remote teaching to allow learning to continue.	Defence	Development of new structures
Video recorded online lessons for the benefit of students who could not attend them (e.g., ill students, students based in a different time zone).		
Developed various strategies to render online teaching more effective (see Tables 4 to 7).		
Implemented synchronous hybrid teaching after schools reopened to mitigate the risk of learning loss for students self-isolating at home.		
Self-isolating teachers delivered lessons remotely from home (where possible), with the support of a teaching assistant who was physically present in the classroom.		
Implemented a rotating three-mode teaching system (i.e., online, in-person, and hybrid lessons) to cope with the learning challenges posed by transport restrictions.		
Surveyed parents and administered tests to diagnose students' learning needs.	Recovery	Exploitation of existing structures
Incorporated more collaborative tasks in in-person lessons to support the recovery of students' social skills.		
Reintroduced guided reading to help strengthen students' reading skills which had overall declined during lockdown.		
Differentiated teaching to provide students of lower academic ability with tailored support.		
Provided catch-up classes during or after school time.		

Wellbeing strategies

As in the case of the learning strategies, the wellbeing strategies employed for defence purposes comprised measures which departed from established practice ("development of new structures"), with the recovery strategies drawing mainly upon existing structures ("exploitation of existing structures"). Interestingly, unlike the learning strategies, the wellbeing ones seemed to be overall fewer in number and less varied. This could reflect a greater readiness on the part of schools to provide learning support compared to mental health aid.

Defence strategies

To prevent students' mental health and overall wellbeing from declining during lockdown, schools employed various defence strategies. These focused mainly on reducing students' screen time. They included measures such as: assigning students non-computer-based tasks; compressing lesson time to allow students a short break away from their computer in between lessons; and increasing the duration of the lunchbreak to encourage students to go outdoors:

> "I would always have something which I called 'Work for the week' which was something which would get the learners away from the computer in the knowledge that in many ways they were spending too much time in front of the screen. So, something that would get them either writing something on paper or reading something away from the computer." [Italy]

> "In the 2021 lockdown, our double lessons, which are usually one hour and 15, were compressed to be only an hour and the idea behind that was to give both students and teachers a little bit of time away from their screens between lessons." [UK]

> "They changed it so that there was a slightly longer break in the middle of the day, so there was longer for lunchtime, to try and encourage people to get outside." [UK]

Teachers also phoned parents and students regularly during lockdown to ensure that any students at risk could be identified as early as possible:

> "During the full lockdown, all pupils were phoned at least once a week, and those pupils who we were particularly worried about were called by members of school staff maybe two or three times a week." [UK]

Recovery strategies

According to the interviewees, schools had a number of mechanisms in place to support students whose mental health and wellbeing were compromised. These typically involved access to a school counsellor or other pastoral support staff, regular one-to-one meetings with teachers, and opportunities for outdoor activities:

> "We've got a school counsellor who's addressing mental health and anxiety issues and things like that. She is available to all the students, and she can be visited on a confidential basis." [Italy]

> "So, there is quite a lot of one-to-one personalised support, and I have several pupils that I meet up with regularly to talk about how everything's going." [UK]

> "This term, we've done a lot of outdoor education to try and build up their wellbeing and that side of things." [UK]

To help teachers increase their knowledge of mental health issues and therefore enable them to support students more effectively, some schools launched mental health training programmes:

> "We are having what's called 'mental first aid training', so any staff that want to volunteer for this training – it hasn't happened yet but it's happening in the future – they'll be trained in mental first aid." [Italy]

A summary of the wellbeing strategies can be found in Table 9.

Table 9: The profile of wellbeing strategies

Micro-level strategies	The "why"	The "how"
Took various measures to reduce students' screen time (e.g., compressed lesson time to allow students a short break away from their computer in between lessons).	Defence	Development of new structures
Phoned students on a regular basis during lockdown to identify any at-risk individuals as early as possible.		
Provided access to a counsellor.	Recovery	Exploitation of existing structures
Provided affected students with regular one-to-one meetings with teachers.		
Provided more opportunities for outdoor activities.		
Launched mental health training programmes to increase teachers' knowledge of mental health issues.		Development of new structures

Discussion

This study attempted to document and understand school responses to the COVID-19 crisis. Through analysing data collected from interviews with teachers based in different parts of Europe, the study identified a number of micro- and macro-level emergency strategies employed by schools to address the challenges posed by the pandemic. As discussed below, apart from providing a useful starting point for any teachers required to deliver emergency education in the future, these strategies also offer valuable insights into schools' adaptability, values and, more importantly, their crisis-readiness. As such, they could prove informative for both educational policy and practice.

What do the strategies reveal about schools' adaptability and values?

The emergency strategies employed by schools were multifaceted: they targeted different areas (*safety*, *learning*, and *wellbeing*), served different functions (*defence* and *recovery*), and employed different problem-solving approaches (*suspension of existing structures*; *exploitation of existing structures*; and *development of new structures*). Overall, these strategies are revealing of schools' agility, adaptability and resilience. Specifically, they are demonstrative of schools' ability to navigate a fast-evolving crisis and respond promptly to challenges, both through exploiting readily available resources and innovating where necessary.

These strategies are also indicative of schools' strong commitment to supporting students in a holistic, equitable and inclusive manner. As the study has shown, schools sought to address not only students' learning needs but also their safety and wellbeing ones. In addition, they aimed to provide all students with the same, or similar, learning opportunities, where possible. For example, they video recorded online lessons to reduce learning loss for students who could not attend them live, and also implemented synchronous hybrid teaching to enable self-isolating students to continue their learning. Furthermore, they strove to provide students – who were affected by the disruption in different ways and to different degrees – with tailored support, through differentiating instruction and providing catch-up classes.

What do the strategies reveal about schools' crisis-readiness?

More importantly, the strategies can provide useful insights into schools' preparedness to cope with a public health crisis similar to that caused by the COVID-19 pandemic. Based on the nature of the strategies employed, various observations can be made about schools' crisis-readiness. Key ones include:

- Overall, the learning strategies identified were considerably more in number and more varied relative to the wellbeing ones. This suggests that the schools participating in the study were not as well prepared to support students' mental health and wellbeing. Given that crises are becoming increasingly more common, this is an area in which the target schools (as well as other schools around the world with similar characteristics) should probably invest more resources to help them become more crisis-ready.

- The recovery strategies identified, that is, the strategies employed to mend, or reverse, the harm caused by the crisis, drew almost exclusively upon existing resources and already established structures ("exploitation of existing structures"). This suggests that the schools in this study have mechanisms in place – albeit probably more learning than wellbeing ones – to support students' recovery in the event of a future emergency.

- Unlike the recovery strategies which capitalised on existing resources and structures, the defence strategies consisted predominantly of attempts to suspend activities ("suspension of existing structures") or devise new solutions to allow the activities to continue ("development of new structures"). This suggests that there were no structures in place which the target schools could exploit or mobilise to defend themselves against the crisis. Overall, the schools seemed to be better prepared to engage in recovery (i.e., to fix the damage caused) than in defence (i.e., to prevent the damage from occurring in the first place), which probably does not represent the most efficient crisis-management approach. To render themselves more crisis-ready and able to respond effectively to another similar public health crisis in the future, the schools may need to invest in developing further their defence capabilities.

Some limitations and directions for further research

When interpreting the findings of the study, two important caveats should be borne in mind. Firstly, given the small scale of the study, the list of strategies reported in this article might not be exhaustive. Secondly, some of the strategies may not be representative of those employed in less affluent contexts, as most of the participants worked in private schools.

Finally, the educational community would benefit considerably from further research into emergency strategies. Such research could focus on capturing strategies employed in a wider range of emergency contexts (e.g., wars, earthquakes, hurricanes) both across the private and state education sectors, as well as on measuring their effectiveness. This would help to extend the present study and support efforts to compile a more comprehensive repository, or database, of emergency strategies which schools around the world can consult whenever they are confronted with a crisis.

Acknowledgements

I would like to thank my colleagues at the Cambridge Centre for Evaluation and Monitoring for their input into the design of the wider study and for their help in recruiting the participants.

References

Acevedo, S., & Novta, N. (2017, November 16). *Climate change will bring more frequent natural disasters and weigh on economic growth*. IMF.

Achtaridou, E., Mason, E., Behailu, A., Stiell, B., Willis, B., & Coldwell, M. (2022). *School recovery strategies: Year 1 findings*. Department for Education.

BERA. (2018). *Ethical guidelines for educational research* (4th ed.). BERA.

Bond, M., Bergdahl, N., Mendizabal-Espinosa, R., Kneale, D., Bolan, F., Hull, P., & Ramadani, F. (2021). *Global emergency remote education in secondary schools during the COVID-19 pandemic: a systematic review*. EPPI Centre, UCL Social Research Institute, University College London.

Braun, V., & Clarke, V. (2021). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology, 18*(3), 328–352.

Carroll, M., & Constantinou, F. (2022). Learning loss in the Covid-19 pandemic: teachers' views on the nature and extent of loss. *Research Matters: A Cambridge University Press & Assessment publication, 34*, 6–25.

Carroll, M. & Constantinou, F. (2023). *Teachers' experiences of teaching during the Covid-19 pandemic*. Cambridge University Press & Assessment.

Constantinou, F. (2023). Synchronous hybrid teaching: how easy is it for schools to implement? *Research Matters: A Cambridge University Press & Assessment publication, 36*, 75–87.

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

Crossfield, J., Behailu, A., Stiell, B., Willis, B., Rutgers, D., Clarkson, L., McCaig, C., Ramaiah, B., & Hallgarten, J. (2023). School recovery strategies: Year 2 findings. Department for Education.

Deng, J., Zhou, F., Hou, W., Heybati, K., Lohit, S., Abbas, U., Silver, Z., Wong, C. Y., Chang, O., Huang, E., Zuo, Q. K., Moskalyk, M., Ramaraju, H. B., & Heybati, S. (2023). Prevalence of mental health symptoms in children and adolescents during the COVID-19 pandemic: A meta-analysis. *Annals of the New York Academy of Sciences, 1520*(1), 53–73.

Di Pietro, G. (2023). The impact of Covid-19 on student achievement: Evidence from a recent meta-analysis. *Educational Research Review, 39*, 100530.

Haileamlak, A. (2022). Pandemics will be more frequent. *Ethiopian Journal of Health Science, 32*(2), 228.

Howard, E., Khan, A., & Lockyer, C. (2021). *Learning during the pandemic: review of research from England*. Ofqual.

Johnson, M. (2022). What are "recovery curricula" and what do they include? A literature review. *Research Matters: A Cambridge University Press & Assessment publication, 34*, 57–75.

Kagawa, F. (2005). Emergency education: a critical review of the field. *Comparative Education, 41*(4), 487–503.

Newton, P. (2021). *Learning during the pandemic: quantifying lost learning.* Ofqual.

OECD. (2022). *Education at a glance 2022: OECD indicators*. OECD Publishing.

Panchal, U., Salazar de Pablo, G., Franco, M., Moreno, C., Parellada, M., Arango, C., & Fusar-Poli, P. (2023). The impact of COVID-19 lockdown on child and adolescent mental health: systematic review. *European Child and Adolescent Psychiatry, 32*(7), 1151–1177.

Senthilingam, M. (2017, April 10). *Seven reasons we're at more risk than ever of a global pandemic.* CNN.

Sharp, C., & Skipp, A. (2022). *Four things we learned about the impact of Covid-19 on mainstream schools and special education settings in 2020 and 2021.* National Foundation for Educational Research.

Sinclair, M. (2001). Education in emergencies. In J. Crisp, C. Talbot, & D. Cipollone (Eds.), *Learning for a future: refugee education in developing countries* (pp. 1–83). UNHCR.

VERBI Software. (2021). MAXQDA 2022 [computer software]. VERBI Software.

World Health Organization. (2020, March 11). *WHO Director-General's opening remarks at the media briefing on COVID-19 – 11 March 2020.*

How long should a high stakes test be?

Tom Benton (Research Division)

Introduction

Educational assessment is used throughout the world for a range of different formative and summative purposes. Wherever an assessment is developed, whether by a teacher creating a quiz for their class, or by a testing company creating a high stakes assessment, it is necessary to decide how long the test should be. Specifically, how many questions should be included and how much time will be required to answer each of them.

The aim of this article is to review some of the most relevant psychometric literature on this topic and show the range of test lengths that would be implied in practice by the various recommendations.

As a counterbalance to this technical work, we also explore the lengths of high stakes assessments across different countries to see how much variation there is. Using international comparisons in this way acts as "a mirror, not as a blueprint" (White, 1987, as cited in Clarke, 2004). What is meant by this is that the lengths of assessments in other countries do not necessarily provide a pattern we should copy. However, by including comparisons to assessment practice in other nations, this research is prevented from becoming purely an exercise in self-justification and we are forced to reflect upon why different countries may come to different conclusions about how long high stakes tests should be.

Before beginning it is worth being clear that, obviously, the answer to the question of how long a test should be will depend upon a range of factors such as its purpose and the breadth of learning it is attempting to assess. Furthermore, the decision requires balancing the costs of long assessments and the impact on the experience of test takers against the likely benefits of increased accuracy. Ultimately such decisions are a matter of educational policy rather than something where a single recommendation can be derived mathematically. Nonetheless, this article attempts to provide practical advice from the perspective of psychometric reliability for considering how long a test should be.

The role of precedent

To start with, it is worth mentioning probably the most influential factor in setting test lengths – the role of precedent.

If a new qualification is intended to be comparable to an existing one, then it would be odd for them to require very different assessment lengths. For example, employers may be sceptical that a qualification requiring only half an hour of assessment will provide the same level of accuracy as one that needed four hours. Conversely, test takers may be upset to be told that the amount of time they need to spend taking exams has been doubled compared to previous years – that is, other test takers have been allowed to achieve the same level of benefit in less time. As such, decisions regarding test length are always likely to build upon what has been done for similar qualifications historically.

Following precedent can also be justified from a technical standpoint. If two qualifications are supposed to be used interchangeably, then it is reasonable to expect that they will measure performance equally accurately. Thus, unless one qualification can achieve high reliability in another way (e.g., adaptive testing), they should be of similar lengths. If reliability differs between two assessments this can have implications for equity. In very broad terms, a short and less reliable assessment will favour less able students as they have an increased chance of overperforming due to good luck. On the other hand, a longer and more reliable test will favour the most able students as it will give them the best chance to demonstrate their skills.

Recommended minimum levels of reliability

Aside from precedent, one way to determine test length is to say that a test should be long enough to meet certain minimum requirements in terms of reliability. Reliability refers to the extent to which we would expect test takers to get the same results were we to replicate the assessment process (Brennan, 2001). For example, the (hypothetical) replication we are interested in might consist of repeating the assessment using different test questions. We would hope that candidates' scores would not change too dramatically if this were done.

Table 1 provides a range of recommended minimum reliability levels for high stakes assessment that can be found in the academic literature. For each of the target reliability values, the second column provides details of at least one of the authors that have suggested it as a minimum. The final column provides some further notes on the language used in relation to this target.

Table 1: A range of minimum reliability levels for high stakes assessment suggested in the academic literature

Target reliability value	Source	Further notes
0.80	Evers (2001), Fry et al. (2012)	"Sufficient", "Typical target"
0.85	Cresswell and Winkley (2012), Frisbie (1988)	"Minimum"
0.90	Evers (2001), Fry et al. (2012), Nunnally (1978) (as cited in Drost, 2011)	"good" or "appropriate" for larger MCQ tests
0.92	Skurnik and Nuttal (1968) and others	Derived from aim that 95 per cent of pupils are accurately classified to within 1 grade. See later discussion in text.
0.95	Kubiszyn and Borich (1993) (as cited in Wright, 1996)	For an "acceptable standardized test"

In interpreting Table 1, it is crucial to note that every author providing these recommendations is clear that reliability will not simply depend upon the characteristics of the test (e.g., its length) but will also be influenced by other factors. To take one example, the quality of the administration conditions may affect the size of reliability coefficients (see Traub & Rowley, 1991, or Frisbie, 1988). Similarly, the authors do not pretend that their suggestions are underpinned by a fully logical argument such as balancing the costs of unreliability against the costs of longer tests. Rather, they simply represent benchmarks based upon the kind of values that have typically been achieved by test developers ever since the easy calculation of reliability indices has been possible.

The target reliability values in Table 1 assume that we are using classical reliability coefficients such as (but not necessarily limited to) Cronbach's alpha (Cronbach, 1951). Such indices of reliability use data on the correlations between scores on items within a test to infer the likely correlation between candidates' observed scores on the test and their scores on another (hypothetical) parallel test.[1] Note that reliability measures of this type are highly dependent upon the ability distribution of the candidates taking them. In particular, they will tend to yield low values in instances where all the students taking a test happen to have very similar levels of ability. To address this concern, the recommendations in Table 1 should be seen as assuming that the range of candidates entering an assessment are broadly representative of the wider population the exam is aimed at. For example, for recommendations to be applicable to a specific GCSE, it should be taken by a similar range of candidates as typically enter GCSEs.

1 A parallel test can be thought of as a test that measures the same constructs as the one being studied, and is equally hard and equally long as the test in question. For example, if two tests fit the Rasch model, they will be parallel if they have identical distributions of item difficulties.

One of the recommended minimum reliability values in Table 1 is 0.92. This is derived from recommendations in the literature relating to classification accuracy. Classification accuracy estimates the percentage of candidates whose grade matches the grade they should be awarded based on their (notional) true score. Their true score is the (hypothetical) score they would achieve on average across many tests parallel to the one they have taken. Classification accuracy is rarely used directly to determine minimum levels of reliability. The reason for this is that, as noted by Wheadon and Stockford (2010), "unless an examination is perfectly reliable, some of those who lie to just one side of a grade will have true scores that fall the other side of it. As a consequence, no examination system can have an accuracy of better than plus or minus one grade" (p. 5). With this in mind, several authors have turned their attention to ensuring that a high percentage of candidates are correctly classified to within plus or minus one grade. Skurnik and Nuttal (1968) suggested a target of ensuring that at least 95 per cent of pupils are accurately classified to within 1 grade. Wheadon and Stockford (2010) agreed that, while this target is essentially arbitrary, it seems a useful point of reference. A similar target (based upon classification consistency) was suggested by Mitchelmore (1981). To convert this suggested target into an equivalent value of classical reliability we have assumed that we are working with the current GCSE grade scale (see footnote for calculation steps[2]).

In summary, Table 1 suggests that, depending upon which author we rely on, the minimum reliability of a test is somewhere between 0.80 and 0.95. Notice that, based on the Spearman-Brown formula (given later) and all else being equal, a test with a reliability of 0.95 will be almost five times as long as one with a reliability of 0.8. Thus, while the exact choice of a target value for reliability may appear to be arguing over tiny details, when it comes to using this to determine test length, a small change can make a big difference.

Having identified a set of recommended minimum reliability levels from the literature, the next step is to estimate how long tests should be to meet these criteria. The steps for this calculation are the subject of the next section.

2 Specifically, from published statistics (GCSE (Full Course) Outcomes for main grade set for each jurisdiction) regarding GCSEs taken in England we know that in summer 2019, 4.5 per cent of candidates achieved grade 9. This implies, if scores were normally distributed, then the grade 9 boundary would be about 1.7 standard deviations above the mean. The same statistics reveal that 98.3 per cent of candidates achieved grade 1 or above meaning that the grade 1 boundary would be 2.1 standard deviations below the mean (if scores were normally distributed). Taken together this means that the eight grade bandwidths (between 1 and 9) would be spread out across 3.8 standard deviations, which in turn implies that the grade bandwidth will be 0.475 standard deviations. For a worst-case scenario of a candidate with a true score directly on a grade boundary, their observed grade will differ from their true grade by more than one if their observed score is too high by *two* grade bandwidths or if it is too low by a single grade bandwidth. This will happen at least 5 per cent of the time if the standard error of measurement is more than 0.28 standard deviations. This indicates a reliability of 0.92 (=1-0.28^2).

Calculating required test lengths

Psychometric formulae

One of the earliest suggested methods for predicting the reliability of a test from its length might be the Spearman-Brown formula (Spearman, 1910; Brown, 1910). This allows us to predict the impact on reliability of lengthening or shortening a test. The Spearman-Brown formula is:

$$\alpha_{comp} = \frac{k\alpha_0}{1 + (k-1)\alpha_0} \quad (1)$$

where α_{comp} is the predicted reliability of a new exam component, α_0 is the known reliability of a reference component, and k is the length of the new exam component relative to the reference one. For example, if we were interested in calculating the likely reliability after doubling the length of a test, k would be set equal to 2.

Similar formulae can be derived starting from an approach to measurement based upon the Rasch partial credit model (Linacre, 2000) so that, under reasonable assumptions, the formula can relate to the total available score in a test and not just the number of items. Other research provides methods to extend the calculations to more complex scenarios such as when combining scores from multiple different assessments potentially measuring different constructs (He, 2009; Wang & Stanley, 1970). In particular, to calculate the reliability of a qualification built from multiple components, all of equal length, and where the separate dimensions of ability they measure are all equally correlated with one another, we can use the following simplification of the Wang-Stanley formula (Wang & Stanley, 1970).

$$\alpha_{qual} = \frac{\alpha_{comp} + (n-1)\rho\alpha_{comp}}{1 + (n-1)\rho\alpha_{comp}} \quad (2)$$

where α_{qual} is the predicted reliability of a new qualification, n denotes the number of components comprising the qualification, and ρ the correlation between true scores in the separate dimensions of ability measured by different components. Note that the formula assumes that all components are equally weighted and that the overall qualification score is obtained simply by adding up all the scores on the components.

The two formulae above can be combined to give:

$$\alpha_{qual} = \frac{k\alpha_0 + (n-1)\rho k\alpha_0}{1 + (k-1)\alpha_0 + (n-1)\rho k\alpha_0} = \frac{k\alpha_0(1 + (n-1)\rho)}{k\alpha_0(1 + (n-1)\rho) + (1 - \alpha_0)} \quad (3)$$

If we want to find the required test length for each qualification component (relative to a known reference component) for a target level of reliability, equation 3 can be rearranged to:

$$k = \left(\frac{\alpha_{qual}}{1-\alpha_{qual}}\right)\left(\frac{1-\alpha_0}{\alpha_0}\right)\left(\frac{1}{1+(n-1)\rho}\right) \quad (4)$$

Finally, we note that our main interest is in the overall length of assessments across the qualification as a whole rather than the length of individual components. That is, we want our formula to suggest values for nk rather than simply k. Putting all this information together yields the following formula for the number of marks to include in a qualification (relative to a reference component with reliability α_0) to achieve a reliability of α_{qual}.

$$nk = \left(\frac{\alpha_{qual}}{1-\alpha_{qual}}\right)\left(\frac{1-\alpha_0}{\alpha_0}\right)\left(\frac{n}{n\rho+(1-\rho)}\right) \quad (5)$$

In order to make use of the above formula, we need values for α_0 and ρ. Ideally, we would like to discuss test length in units of time (e.g., minutes) rather than in terms of the number of available marks. For this reason, we also need to know how many minutes are typically allowed for each available mark in an exam. All of these matters are discussed next.

Reliability of reference component

First, we attempt to identify a suitable value for α_0. This can be done by looking at empirical data on test reliability historically.

By far the largest amount of published data on test reliability is in the form of Cronbach's alpha. This type of data provides a natural starting point for calculations. For example, Bramley and Dhawan (2010) published a wealth of information on the reliability of OCR examinations such as a chart showing how Cronbach's alpha increases along with the number of marks in a test (see their Figure 1.4). A similar chart, based on all OCR GCSE and AS/A Level components (that is, individual examination papers) taken by at least 500 candidates across the five years from the start of 2015 until the end of 2019, is shown in Figure 1.[3] This chart summarises the reliability coefficients associated with almost 1600 assessments. Assessments are grouped by rounding the number of available marks to the nearest 10, and the distribution of reliabilities within each group is shown in the form of a boxplot. The largest number of assessments (more than 300) had a maximum mark of 60. As can be seen, for this maximum mark band, the reliability coefficients were just above 0.8 on average. Slightly fewer assessments (but still more than 200) had a maximum mark of 50. The average reliability for these assessments was very close to 0.8. Very few assessments had maximum marks below 50 and so these elements of the chart can be ignored.

3 Figure 1 also includes reliability estimates for papers with optional questions. In these cases, Backhouse's formula P (Backhouse, 1972) is used as a substitute for Cronbach's alpha.

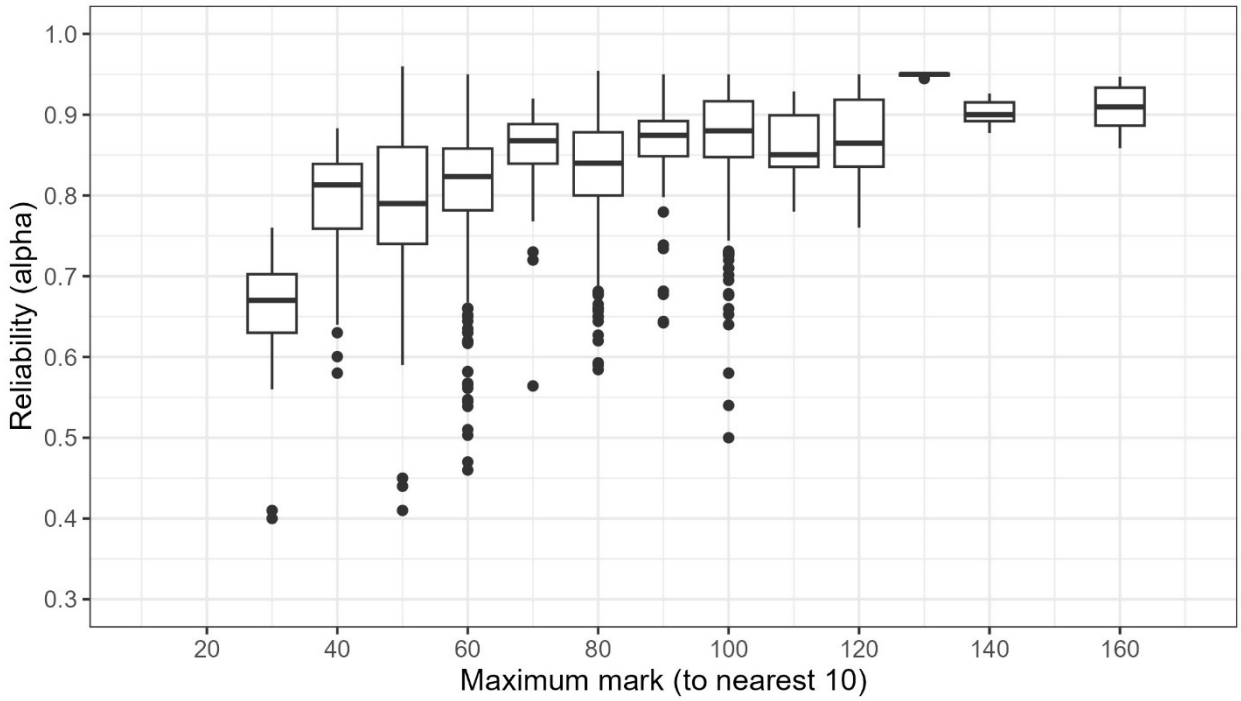

Figure 1: Relationship between test reliability and total test mark for all OCR GCSE and AS/A Level examinations entered by at least 500 candidates between 2015 and 2019

If we assume that an average test with 50 marks has a reliability of 0.8, then according to the Spearman-Brown formula, tests with maxima of 60, 70, 80 and 90 should have reliabilities of 0.83, 0.85, 0.86 and 0.88. Broadly speaking, this fits with the average reliabilities we can see in Figure 1.

However, if we continue to use the Spearman-Brown formula, we would expect tests with maxima of 100, 120 and 140 to have reliabilities of 0.89, 0.91 and 0.92 respectively. These expectations are not reflected by the data in Figure 1. This is likely to be because, as mentioned above, reliability coefficients depend upon a range of factors that may be associated with test length and not just test length itself. In particular, in our data, longer tests are more likely to be part of an A Level, and shorter ones more likely to be part of a GCSE. A Levels tend to have slightly lower reliability coefficients for the same test length (perhaps due to the more restricted range of candidates involved). For example, among 60 mark tests the median reliability of a GCSE component is 0.83 whereas for an A Level it is 0.81.

Despite the differences between qualifications, it seems reasonable to use the starting point of 0.8 for a 50-mark test because calculations of test length should evaluate how reliability changes with test length within a fixed group of candidates.

Compared to some published statistics of test reliability, a starting point of 0.8 for a 50-mark test may seem disappointingly low. For example, recent published statistics for Key Stage 1 national curriculum tests in English suggest that these 40-mark reading tests for 7-year-olds achieve a reliability of about 0.95.[4] However,

4 See Tab 28 of National Curriculum Test Handbook: 2016 and 2017 technical appendix.

the apparently high reliability in that context may be because of the very large diversity in reading ability among children of that age. As such, it may not be something we would expect to repeat at GCSE. This is partially confirmed by the fact that the same set of published statistics suggest that the reliability of 50-mark national curriculum reading tests for 11-year-olds (Key Stage 2) is lower at 0.89.

Note that, in using this starting point we are not assuming that every 50-mark test will always yield an alpha coefficient of 0.8. The exact values of reliability coefficients are dependent upon numerous factors. In particular, the range of abilities of the candidates taking the test will have a big influence. However, this factor is largely beyond the control of the test developer. What we can do is try to create a test with sufficient length such that, assuming it were taken by a set of candidates with a range of abilities typical of those entering a GCSE, we would have a good chance of alpha exceeding some target value. The starting assumption that a 50-mark test will typically have a reliability of 0.8 (under these circumstances) allows us to do exactly that. To put it another way, for the purposes of using our formula we will set α_0 to be 0.8 and assume that this refers to a reference test form with a maximum mark of 50.

Correlation between true scores on different components

In order to apply equation 5, we also need a value for the correlation between true scores on different components (ρ). Such a value can be obtained using information in Benton (2021a) which indicates that the correlation between observed scores on separate components within an A Level is typically 0.64 while the median reliability of the same components is 0.83. Combining this formula with Charles Spearman's 1904 correction for attenuation formula (Spearman, 1987) yields a value of just below 0.8 ($\approx 0.64/\sqrt{0.83*0.83}$) for the estimated correlation of true scores on separate components. We will use this value in our calculations of required test lengths.

Note that performing the same calculations based on GCSEs taken in summer 2019 leads to a somewhat higher value for the correlation (approximately 0.9). However, as we will see, even with a value of 0.8, accounting for qualifications consisting of multiple components (presumably measuring slightly different skills on different occasions) has a fairly limited impact on the amount of assessment time required in total.

Time per mark

Finally, we require a clear understanding of the usual relationship between the maximum available mark on a test and its (usual[5]) duration in minutes.

5 In England, exam candidates with special educational needs, disabilities or temporary injuries can be allowed extra time to complete an examination if they need it. For the purposes of this paper, we focus on the amount of time that is allowed to students without these access arrangements.

The relationship between the number of marks in an examination and duration is shown in Figure 2. Each point in the chart represents an OCR exam component taken between 2015 and 2019. Separate chart panels have been used for GCSEs and AS/A Levels. A small amount of jitter has been added to the points in the chart to allow the distribution of times and total marks to be seen more clearly. The dashed diagonal lines represent lines of equality. A blue regression line, based upon regression through the origin, is also included. Regression through the origin was used as it is consistent with the (sensible) idea that an exam with no marks would be expected to take no time.

Across both qualification types, the number of minutes allowed for an exam is rarely less than the total number of marks and is usually slightly higher. This fits with the idea in assessment folklore of "a mark a minute" – although internet searches suggest this phrase is used far more often as a guide for students about how long they should spend on exam questions rather than for test developers deciding upon exam duration. The gap between test length in marks and duration in minutes is slightly larger for A Levels than for GCSEs.

Based upon the regression lines, in broad terms, the number of minutes allowed for an exam has tended to exceed the number of available marks by about 20 per cent. We will use this figure as a basis to identify the likely duration of tests needed to meet the reliability thresholds listed earlier.

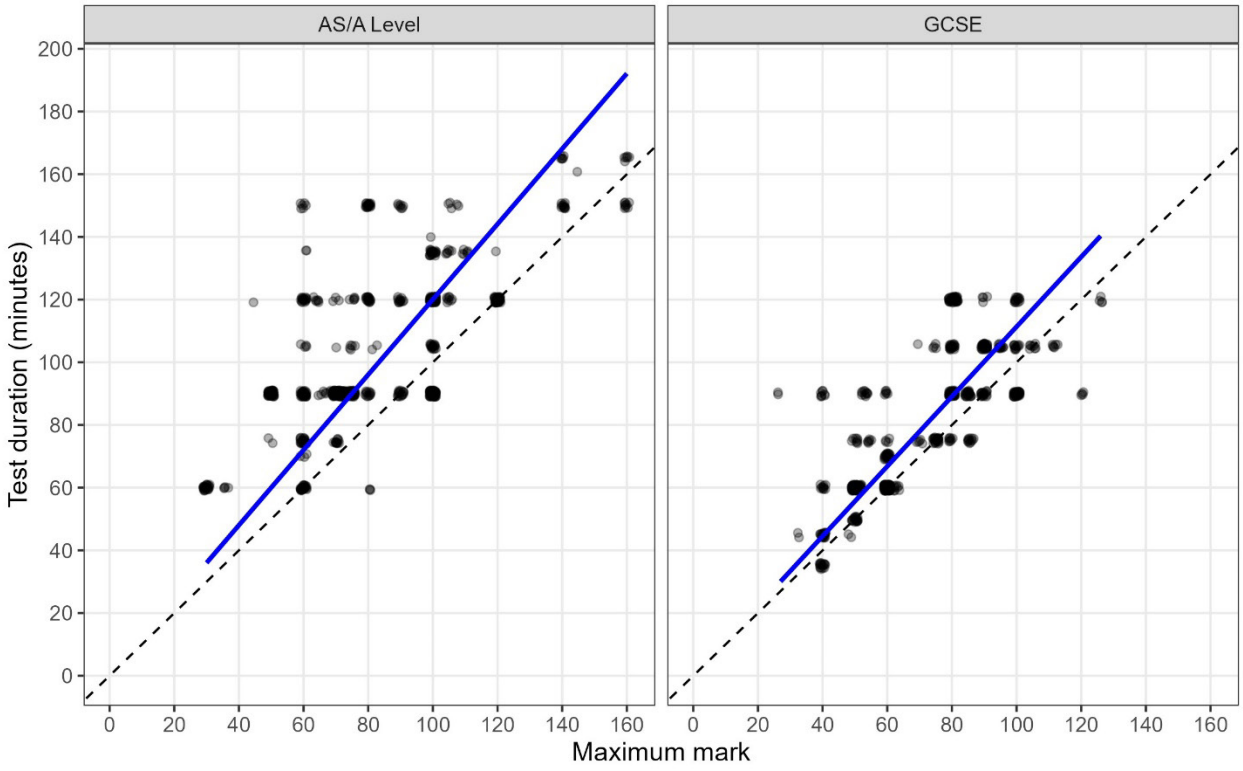

Figure 2: Relationship between test length and test duration for all OCR GCSE and AS/A Level examinations entered by at least 500 candidates between 2015 and 2019. A solid blue regression line (regression through the origin) is included within each chart. The dashed line represents a line of equality between test length in marks and duration in minutes.

Suggested test durations

By combining all of the above assumptions with the formula in equation 5, we can derive the following formula for the total amount of assessment time that is required for a qualification with n components to likely achieve a reliability α_{qual}.

$$Total\ required\ assessment\ time = 15\left(\frac{\alpha_{qual}}{1-\alpha_{qual}}\right)\left(\frac{n}{0.8n+0.2}\right) \quad (6)$$

This formula is derived from equation 5 but replacing both α_0 and ρ with 0.8, multiplying by 50 as our reference reliability comes from a test with 50 marks, and multiplying by 1.2 as we (currently) typically allow 1.2 minutes for each available mark in a test.

Using equation 6, Table 2 shows how the recommended length of tests varies according to the target for reliability we use to determine test length, and the number of components of which the qualification is comprised. The lowest target for reliability considered in this table is the value of 0.80 from Evers (2001). To hit this benchmark our analysis suggests that a high stakes qualification should comprise of at least 50 marks and require about an hour of exam time at a minimum. If the qualification comprises of several components, presumably measuring different skills on different occasions, then the total exam time should increase by perhaps 10 minutes. In other words, spreading measurement across different components has only a minor impact on the total amount of assessment time required to meet reliability requirements.

As expected, as reliability requirements become more stringent, the suggested test lengths increase. Aiming for a reliability coefficient of 0.9 requires a total exam time of between 2 and 3 hours. Aiming for Wheadon and Stockford's (2010) point of reference that qualifications should classify students into the correct grade plus or minus one at least 95 per cent of the time (i.e., a reliability of 0.92) generally requires total examination times in excess of 3 hours. Finally, to achieve the most stringent reliability target we have considered (0.95) will typically require between 5 and 6 hours of assessment.

Table 2: Estimated required total minutes of assessment depending upon target reliability level and the number of components in the assessment

Target reliability	Number of assessment components				
	1	2	3	4	10
0.8	60	67	69	71	73
0.85	85	94	98	100	104
0.9	135	150	156	159	165
0.92	173	192	199	203	210
0.95	285	317	329	335	348

At this point, those involved in the creation and regulation of GCSEs in England may be tempted to congratulate themselves. As it happens, a typical GCSE in England consists of two components (i.e., two separate exam papers) and requires roughly 3 and a half hours of exam time in total. Based on Table 2, this is only slightly higher than the recommended amount of assessment (192 minutes) needed to achieve a qualification reliability of 0.92. From our discussion earlier, this is also roughly the test length required to ensure that, 95 per cent of the time, the grades awarded to candidates are equal to their true grades plus or minus one. However, to avoid this article descending into self-congratulation, and to force us to reflect more deeply on the length of assessment that is actually needed at different stages of education, we next compare the amount of time spent in high stakes examinations in England to that in other countries.

Test lengths in high-performing jurisdictions

Table 3 provides a summary of test durations for qualifications taken in England as well as qualifications/assessments taken in 10 high-performing jurisdictions. The 10 comparator jurisdictions in this article have been chosen from those identified in Elliott et al. (2015) and Suto and Oates (2021). Only assessments that are high stakes for the pupil (leading to a recognised qualification) or are compulsory within their region are included. In addition, the focus is on assessments taken at similar ages to GCSEs and A Levels. For example, although the NAPLAN tests are taken in grade 6 (age 11/12) and grade 9 (age 14/15) in Victoria, the details in the table are based on the grade 9 tests. Note that not all countries identified in Elliott et al. (2015) and Suto and Oates (2021) are included here. This is due to not finding detailed information on the duration of examinations in some countries at the time of undertaking the review for this research in early 2021. Nonetheless, although Table 3 is far from being a comprehensive review of the durations of compulsory and high stakes examinations in high-performing jurisdictions, it hopefully provides a sufficiently wide variety of decisions to facilitate further discussion about test lengths. Links to the websites that were used as a source of information are provided at the end of the article.

As shown in Table 3, and based on qualifications awarded in summer 2019, GCSEs in England require an average of 3.5 hours of exams[6] (typically two exams of an hour and 45 minutes each), whereas A Levels require 6 hours on average[7] (typically three exams of 2 hours). As such, both qualifications are long enough to generally meet some of the highest benchmarks for reliability displayed earlier in Table 2.

Exams at ages 14 to 17

Table 3 allows us to compare the duration of GCSEs to the duration of other exams taken by students of similar ages in education systems around the world.

6 Excluding double science (which counts as two qualifications) and restricting to GCSEs that currently use exams only for assessment.

7 Also restricted to A Levels assessed using exams only. The A Levels requiring the longest exam time are Latin and Classical Greek (7 hours each). All others with these criteria require 6 hours of exam time.

As can be seen, the majority of such assessments require considerably less time per subject than GCSEs. The shortest such assessments are the NAPLAN tests in Australia (Victoria) where the longest assessments (reading and numeracy) take only slightly over an hour each. The relatively short duration of the NAPLAN assessments might be justified by the fact that, although compulsory, the exams are relatively low stakes with the central purpose being to monitor student progress. Similar comments might be made about the assessments used within the Provincial Achievement Testing Program in Canada (Alberta), many of which only require a little over an hour each.

Junior Leaving Certificate exams in the Republic of Ireland have slightly higher stakes as they form part of graduation from secondary school. This may be reflected in the slightly longer required amount of exam time per subject (2 hours). Note that, for these qualifications, marks from exams are supplemented by an additional 10 per cent of marks that are available via school-based assessments. Required exam times for exams taken at ages 14 to 17 in New Zealand (3 hours), Singapore (3.5 hours), and Massachusetts (4 hours) are more similar to those required for GCSEs in England. However, to set this comparison in context we need to consider how many subjects students enter on average. In England, students take nine GCSEs on average (Carroll & Gill, 2018). As such, we expect the average GCSE student in England to spend almost 32 hours taking exams. In contrast, in Singapore the maximum (not the average) number of O Levels a student can take is nine (in the Special and Express stream), and most students will take fewer than this. The maximum number that can be taken in the Normal (Academic) stream is seven. Similarly, according to UCAS, in New Zealand students are typically required to study between five and six subjects for each level of NCEA. In Massachusetts, graduation only requires that students pass exams in three subjects. As such, the total amount of time spent in exam rooms will be substantially lower in these jurisdictions than for students taking GCSEs in England.

The Comprehensive Assessment Programme (CAP) in Chinese Taipei provides an interesting alternative set of arrangements to the GCSE. It is taken at a similar age to GCSEs and is high stakes in that it is a required part of progression to the next stage of education. It relies entirely on external assessment in the form of examinations. However, rather than requiring lengthy separate examinations for different subjects, all subjects are assessed in 7 hours of assessments split across two days. This represents an intense assessment procedure for the student but one that requires far less time than is needed for a student in England to complete all of their GCSEs. In fact, considered as a whole, the CAP actually represents one of the shortest total assessment times of any of the high stakes exams at age 14–17 shown in Table 3. The reasons why shorter assessment time is possible for CAP are not clear. However, it would seem likely that a focus on overall achievement across all subjects rather than a need to have highly reliable assessment for each individual subject may partially explain the difference.

End of secondary and university entrance

From Table 3, the total amount of examination time required for A Levels in England does not seem unusual compared to other countries that focus on individual subjects. For example, both New Zealand (3 hours) and Poland (4 hours) tend to require slightly less examination time per subject but will also typically require students to study larger numbers of subjects (five or six in New Zealand, at least four in Poland). Note that the NCEA in New Zealand also incorporates a substantial amount of internal assessment. The amount of exam time per subject in Canada (Alberta) is the same as A Levels in England. However, it is worth noting that, in contrast to England, despite their length, Diploma Examinations in Alberta only provide 30 per cent of each student's final qualification mark with the remainder dependent upon schools' own assessments.

An interesting contrast to the amount of time required for A Levels is provided by university entrance exams in Japan and South Korea. These exams are extremely high stakes for students as they are the primary means of determining university entrance. However, as a whole they require considerably less exam time than in A Levels in England. Whereas students in England are typically required to spend between 18 and 24 hours taking exams (depending upon the number of A Levels they study), in Japan all assessment is completed in 12 hours (spread over two days) and in South Korea it is completed in 6.5 hours (all on one day). The reduced total assessment time may be because of the very clear single purpose of the exams (university entrance) and the resulting possibility of focusing on results across all subjects combined rather than needing highly reliable results in each individual subject. Of course, the highly compressed timescale for assessment in these countries (one or two days) also has some disadvantages such as the amount of pressure it places on students.

Table 3: Times required for various examinations in England and other high-performing jurisdictions

Country	Assessment name	Target group	Typical exam time required	Additional internal assessment	Number of subjects taken
Australia (Victoria)	National Assessment Program – Literacy and Numeracy Testing (NAPLAN)	Year 9 (Age 14/15)	40–65 minutes per subject	No	4
Canada (Alberta)	Diploma Examination	End of senior high school (university entrance)	Up to 6 hours per subject[8]	Yes. 70% of the final course-mark is derived from internal assessment.	Unclear
Canada (Alberta)	Provincial Achievement Testing Program	Grade 9 (Age 14/15)	1.25 to 3.25 hours per subject[8]	No	4
Chinese Taipei	Comprehensive Assessment Programme for Junior High School Students (CAP)	Year 9 (Age 14/15)	7 hours in total	No	5
England	GCSE	Year 11 (Age 15/16)	3.5 hours per subject	Only in a minority of subjects	9 on average
England	A Level	End of secondary education	6 hours per subject	Only in a minority of subjects	3 or 4
Japan	National Center Test for University Admissions	University entrance	12 hours (approx.) in total	No	6 (if separate sciences counted as one subject each)
New Zealand	National Certificate of Educational Achievement (NCEA)	Year 11 (Age 15/16) through to end of secondary education	3 hours per subject (all levels)	Yes. Internal and external assessments both feed into a credits system.	Typically 5 or 6
Poland	egzamin maturalny ("Matura")	End of secondary education	3 hours per subject	No	At least 4
Republic of Ireland	Junior Certificate	Third year of Junior Cycle (Age 15/16)	2 hours per subject	Yes. 10% of qualification marks from internal assessment.	Possibly 7 or 8 per pupil on average[9]
Singapore	O Levels	Secondary years 4 or 5 (Age 14–17)	3.5 hours per subject	No	Between 4 and 9
South Korea	College Scholastic Ability Test (CSAT)	University entrance	6.5 hours in total	No	7
USA (Massachusetts)	Massachusetts Comprehensive Assessment System (MCAS)	Grade 10 (Age 15/16)	Untimed but recommended time is 2 hours per subject	No	At least 3

8 Intended time for students to complete the test. Since 2017 all students are allowed double this amount if they desire it.
9 Based on dividing the total number of entries to Junior Certificate exams by an estimate of the number of eligible pupils.

Summary and discussion

Reflecting on the results in this paper we can see that, although psychometrics can help us think about how long exams need to be to achieve acceptable reliability, ultimately, the decision is one of policy. The costs of increasing the length of exams in terms of the burden on students, schools and assessors, need to be balanced against the likely benefits of reliable assessment such as public confidence and ensuring that the correct decisions are made about students' futures.

The need for judgement in making this decision can be seen in several ways throughout this article. Firstly, while psychometrics has supplied formulae relating test length to reliability, different authors have made different recommendations regarding what level of reliability is acceptable. Secondly, a brief review of test lengths from different countries around the world reveals a fairly wide variety of approaches in practice.

It is clear that GCSEs and A Levels in England are of sufficient length to likely meet the levels of reliability that are recommended in the academic literature. However, some of the (less stringent) recommendations might also be met by somewhat shorter examinations. Furthermore, comparison with decisions in other countries make it clear that different decisions are possible. This is particularly evident for examinations taken at ages 14–17 where the total exam time for GCSEs in England appears relatively high compared to other countries.

To some extent, differences in length can be explained by differences in purpose. In particular, some of the shortest examination lengths were seen for assessments that are primarily formative in their purpose such as NAPLAN in Australia (Victoria) or the Provincial Achievement Testing Program in Canada (Alberta). Nonetheless, there are also examples of countries such as the Republic of Ireland where high stakes qualifications are awarded based on substantially shorter exams than in England. Furthermore, although O Level exams in Singapore are of similar length to GCSEs in England, students tend to take fewer such exams.

From the analysis of the length of exams in other jurisdictions (e.g., CAP tests in Chinese Taipei; university entrance tests in South Korea and Japan), it seems possible to reduce the total exam length by focusing on overall achievement across all subjects, rather than attempting to provide highly reliable assessment for each one individually.

Decisions about test length require a clear understanding of the purposes of assessment. This would certainly include considering whether an assessment is primarily formative or summative as well as how it may be combined with other information to impact on decisions about students' futures. It might also include a consideration of comparability and ensuring that any new qualification meets broadly the same requirements as existing ones to which it will be compared.

Limitations

The calculations in this report have been derived from looking at average reliabilities across lots of assessments. As such, while they are intended to provide a reasonable guideline to help in determining test lengths, they may be less appropriate in situations where we have good reasons to expect reliability to differ from the typical situation. One example might be where we expect a greater amount of variation between markers. Since, all else being equal, marking error will tend to have some impact on reliability coefficients such as alpha, we may reasonably expect exams consisting of a few essays to have lower reliabilities than suggested by the formulae in this article. As such, we may wish to compensate by increasing the number or length of such exams in a qualification. A more detailed consideration of the relationship between marking error, reliability, validity and recommended test lengths could be the subject of further research.

Recommendations

As is clear from the above discussion, there is no single correct answer to the question of how long a test should be. However, there are perhaps a few general principles that are always worthy of consideration in making this decision. Based on the research described in this article, some potential principles are:

- If the purpose of a test is primarily to provide formative feedback to a student on their progress, a test length of about one hour would be fairly typical of what is required in different countries.
- If an assessment is expected to have a direct impact, on its own, on decisions made about individual students then, for consistency with all but the most permissive psychometric criteria, the test should be at least 90 minutes long. Having said this, there are a few possible justifications for shorter assessments:
  - If they are measuring a very narrow construct. For example, a test of whether primary school children know their times tables, or whether they can read words using phonics, could not reasonably be expected to take longer than half an hour for each student.
  - If computer adaptive testing is used to achieve reliable assessment in a shorter amount of time (but see Benton, 2021b, for a wider discussion of this).
- If the primary focus of assessment is on overall performance across subjects (rather than within each individual subject), as little as one hour per subject may be sufficient to achieve reasonable reliability.
- If an assessment is for students' final qualifications before university, at least 3 hours of exam time per subject is not unusual internationally. Given the high stakes of qualifications taken at this age, this would appear to be a sensible lower bound for test length.
- If a new qualification needs to be directly comparable to an existing one (e.g., for use in school performance tables), it is sensible to ensure that elements of assessment design such as test length are kept reasonably similar.

References

Backhouse, J. K. (1972). Appendix two: Mathematical Derivations of Formulae P, Q, Q', and S. In Nuttall, D. L., & Willmott, A. S. (1972). *British examinations: techniques of analysis*. National Foundation for Educational Research in England and Wales.

Benton, T. (2021a). On using generosity to combat unreliability. *Research Matters: A Cambridge Assessment publication, 31*, 22–41.

Benton, T. (2021b). Item response theory, computer adaptive testing and the risk of self-deception. *Research Matters: A Cambridge University Press & Assessment publication*, *32*, 82–100.

Bramley, T., & Dhawan. V. (2010). *Estimates of reliability of qualifications*. Ofqual/11/4826.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, *38*, 295–317.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3,* 296–322.

Carroll, M., & Gill, T. (2018). *Uptake of GCSE subjects 2017*. Statistics Report Series No. 120. Cambridge Assessment.

Clarke, D. (2004). Researching classroom learning and learning classroom research. *The Mathematics Educator*, *14*(2).

Cresswell, M., & Winkley, J. (2012). *Introduction to the concept of reliability*. Chap. 1 in *Ofqual's Reliability Compendium*, edited by D. Opposs and Q. He. Ofqual/12/5117.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Drost, E. A. (2011). Validity and reliability in social science research. *Education Research and Perspectives, 38*(1), 105–123.

Elliott, G., Rushton, N., Darlington, E., & Child, S. (2015). Are claims that the GCSE is a white elephant red herrings? Cambridge Assessment Research Report.

Evers, A. (2001). Improving test quality in the Netherlands: Results of 18 years of test ratings. *International Journal of Testing*, *1*(2), 137–153.

Frisbie, D. A. (1988). NCME Instructional Module on reliability of scores from teacher-made tests. *Educational Measurement: Issues and Practice*, *7*(1), 25–35.

Fry, E., Crewe, J., & Wakeford, R. (2012). The Qualified Lawyers Transfer Scheme: innovative assessment methodology and practice in a high-stakes professional exam. *The Law Teacher*, *46*(2), 132–145.

He, Q. (2009). *Estimating the reliability of composite scores*. Ofqual/10/4703.

Kubiszyn, T., & Borich, G. (1993). *Educational testing and measurement*. Harper Collins.

Linacre, J. M. (2000). Predicting reliabilities and separations of different length tests. *Rasch Measurement Transactions*, *14*(3), 767.

Mitchelmore, M. C. (1981). Reporting student achievement: How many grades? *British Journal of Educational Psychology, 51*(2), 218–227.

Nunnally, J. C. (1978). *Psychometric theory*. McGraw-Hill Book Company, pp. 86–113, 190–255.

Skurnik, L. S., & Nuttal, D. L. (1968). Describing the reliability of examinations. *Journal of the Royal Statistical Society. Series D (The Statistician), 18*(2), 119–128.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3,* 271–295.

Spearman, C. (1987). The proof and measurement of association between two things. *The American Journal of Psychology, 100*, 441–471.

Suto, I., & Oates, T. (2021). *High-stakes testing after basic secondary education: How and why is it done in high-performing education systems?* Cambridge Assessment Research Report.

Traub, R. E., & Rowley, G. L. (1991). An NCME instructional module on understanding reliability. *Educational Measurement: Issues and Practice*, *10*(1), 37–45.

Wang, M., & Stanley, J. (1970). Differential weighting: A review of methods and empirical studies, *Review of Educational Research*, *40*, 663–705.

Wheadon, C., & Stockford, I. (2010). *Classification accuracy and consistency in GCSE and A Level examinations offered by the Assessment and Qualifications Alliance (AQA) November 2008 to June 2009*. Ofqual/11/4823.

White, M. (1987). *The Japanese educational challenge: A commitment to children*. The Free Press.

Wright, B. D. (1996). Reliability and separation. *Rasch Measurement Transactions*, 9(4), 472.

Sources for lengths of assessments in different jurisdictions

Links to the websites that were used as a source of the information in Table 3 are listed below. All websites were last accessed by the author on 25 April 2024.

Australia (Victoria)
https://www.vcaa.vic.edu.au/Documents/naplan/parentpamphlet/2021/NAPLANonPaperInformationforParentsandCarersBrochure.pdf

Canada (Alberta)
https://www.alberta.ca/education-guide-diploma-examinations-program (Diploma examinations).
https://www.alberta.ca/system/files/custom_downloaded_images/ed-diploma-exam-general-information-bulletin.pdf (Diploma examination schedules)
https://www.alberta.ca/system/files/custom_downloaded_images/edc-pat-general-information-bulletin.pdf (Provincial Assessment Tests)

Chinese Taipei
https://en.wikipedia.org/wiki/Comprehensive_Assessment_Program_for_Junior_High_School_Students
https://ncee.org/country/taiwan/

England
Test lengths are published by each awarding organisation and can be found by searching for the "specifications at a glance" for a particular subject at either GCSE or A Level. Some example links are provided.
https://www.aqa.org.uk/subjects/science/gcse/biology-8461/specification-at-a-glance
https://qualifications.pearson.com/content/dam/pdf/GCSE/mathematics/2015/specification-and-sample-assesment/gcse-maths-2015-specification.pdf
https://www.ocr.org.uk/qualifications/as-and-a-level/economics-h060-h460-from-2019/specification-at-a-glance/

Japan
https://en.wikipedia.org/wiki/National_Center_Test_for_University_Admissions
https://www.japaneducation.info/tests/higher-education-tests/national-centre-test-for-university-admissions.html

New Zealand
https://www2.nzqa.govt.nz/assets/About-us/Publications/Resources-and-videos/Guide-to-NCEA/Guide-to-NCEA-English.pdf
https://www2.nzqa.govt.nz/assets/NCEA/2024-Timetable-FINAL-4-3-24.pdf
https://qips.ucas.com/qip/new-zealand-national-certificate-of-educational-achievement-level-3-ncea-level-3

Poland
https://en.wikipedia.org/wiki/Matura#In_Poland
https://cke.gov.pl/images/_KOMUNIKATY/20201222%20E8%20EM%20Komunikat%20o%20harmonogramie%20AKTUALIZACJA.pdf (Translated using Google Translate)

Republic of Ireland

Examples of links to assessment details for Junior Certificates in individual subjects are provided.

https://www.jct.ie/perch/resources/maths/junior-cycle-mathematics-specification-2018.pdf

https://www.curriculumonline.ie/getmedia/2a7a8d03-00e6-4980-bf20-f58def95688f/JC_Geography-en.pdf

https://www.curriculumonline.ie/junior-cycle/junior-cycle-subjects/modern-foreign-languages/assessment-and-reporting/

Singapore

https://en.wikipedia.org/wiki/Singapore-Cambridge_GCE_Ordinary_Level.

Details can also be found in syllabus descriptions, as in these examples.

https://www.seab.gov.sg/docs/default-source/national-examinations/syllabus/olevel/2022syllabus/2273_y22_sy.pdf

https://www.seab.gov.sg/docs/default-source/national-examinations/syllabus/olevel/2022syllabus/7087_y22_sy.pdf

https://www.seab.gov.sg/docs/default-source/national-examinations/syllabus/olevel/2023syllabus/4052_y23_sy.pdf

South Korea

https://en.wikipedia.org/wiki/College_Scholastic_Ability_Test

USA (Massachusetts)

https://www.doe.mass.edu/mcas/graduation.html (Number of required subjects)

https://www.doe.mass.edu/mcas/testadmin/manual/tam-cbt.pdf (Recommended session durations)

Test design documents provide information on the number of sessions required for each subject for the Grade 10 tests. Some examples are given.

https://www.doe.mass.edu/mcas/tdd/ela.html?section=testdesign

https://www.doe.mass.edu/mcas/tdd/math.html?section=testdesign

https://www.doe.mass.edu/mcas/tdd/sci.html?section=testdesign

Core Maths: Who takes it, what do they take it with, and does it improve performance in other subjects?

Tim Gill (Research Division)

Introduction

Core Maths (CM) qualifications were introduced into the post-16 curriculum in England in 2014, with first assessments in 2016. They are a suite of qualifications aimed at students who achieve a pass grade (grade 4) or higher at GCSE Maths (taken at age 16) but do not go on to take AS or A Level Maths (at age 17 or 18). This group comprised around 40 per cent of all 16-year-old students in 2013, when the qualification was proposed (DfE, 2013). The main purposes of introducing CM were to increase participation in post-16 maths, and to help develop students' mathematical knowledge and its application to a range of different areas. This means these qualifications may help students in subjects which have some mathematical content, such as psychology, business, engineering, and sciences. CM qualifications also have a focus on the application of mathematical techniques to real-world contexts.

There are several different qualifications currently within the CM suite, offered by different awarding organisations (AOs). Some AOs offer more than one CM qualification, each with a different focus. For example, OCR (Oxford, Cambridge & RSA) currently offers two CM specifications (Core Maths A and Core Maths B) and provides some guidance on its website[1] as to which specification to choose, based on the content and what other subjects are supported:

> "Core Maths A content supports all Level 3 qualifications which have a quantitative skills requirement. This includes, but is not limited to: business and economics, PE [physical education] and sport, health and social care, design and technology, engineering and all the science subjects.
>
> Core Maths B content supports subjects that require statistical skills, such as biology and environmental science, psychology, geography and sociology."

1 https://www.ocr.org.uk/qualifications/core-maths/

The qualifications are designed to be taken over two years and are equivalent to half an A Level. However, there is evidence that some schools offer it as a one-year course (Homer et al., 2020).

There is limited previous research into whether the qualifications' aims have been achieved. Homer et al. (2020) undertook a review of the qualification in its "early years" (2016 to 2019), including analysis of the characteristics of students taking CM qualifications, what other qualifications and subjects were taken alongside, and whether there was evidence that CM students performed any better than non-CM students in A Levels with some numeric content. In terms of the student characteristics, they found that the percentage of female students increased from 34 per cent in 2016 to 45 per cent in 2019, and that in 2019 CM students were, on average, more deprived than students taking A Level Maths, but less deprived than students not taking any Key Stage 5 (KS5) maths qualification. In 2018, the most common subjects taken alongside CM were mostly popular AS or A Levels with a quantitative element (e.g., Maths, Psychology, Business Studies, Chemistry) and the Extended Project Qualification (EPQ). They found no evidence that taking CM was associated with better performance in selected A Levels taken at the same time (even after accounting for other factors including prior attainment, gender, deprivation, and school type).

Homer et al. (2020) also surveyed teachers and students to elicit views of the qualification. Both groups tended to be positive about it, particularly its applications to real-world situations. They also believed that CM supported students in their other subjects with mathematical content taken concurrently, although this belief was not backed up with any empirical evidence of improved performance, as already discussed.

Uptake of CM qualifications has increased since its introduction, from 2930 in 2016 to 12 367 in 2023 (AMSP, no date). However, this is still some way below expectations. According to the Royal Society (2023), entries in 2021/22 amounted to just 7 per cent of the potential candidates (i.e., those taking A Levels, but not AS or A Level Maths). This demonstrates that one aim of the qualification (to significantly increase uptake of maths post-16) has not been achieved. Their research also found that provision of CM throughout England was "patchy", with the proportion of schools and colleges offering the subject varying greatly between different local authorities. They called for more recognition from universities, such as inclusion of the qualification in entry requirements for students. It is worth noting that some universities already recognise the benefits of CM and make alternative offers to students taking it[2].

Since the investigation of the impacts of CM in its "early years", as described in Homer et al. (2020), there has been no more recent evaluation of its possible benefits. The research presented here aimed to bring up to date some of this previous analysis. The main purpose was to investigate whether there is any evidence that taking a CM qualification is beneficial to students in terms of their performance in other qualifications taken concurrently (e.g., A Levels, BTECs,

2 See https://amsp.org.uk/universities/university-admissions/alternatives-admissions/

or Cambridge Technicals). This analysis was restricted to subjects with some quantitative element, as these were the subjects that the qualifications were meant to support and, therefore, the most likely area of benefit.

We also investigated the background characteristics of students taking CM, and which other qualifications and subjects CM was most likely to be combined with. In particular, we investigated if there have been changes in uptake since the work of Homer et al. (2020), expanded on their analysis to include more student and school characteristics, and carried out a more in-depth look at the qualifications and subjects combined with CM.

The research questions were:

I. What are the background characteristics of Core Maths students (e.g., gender, prior attainment, ethnicity)?
2. Which other qualifications (e.g., A Levels, BTECs, Cambridge Technicals) and subjects are most likely to be taken alongside Core Maths?
3. Does Core Maths provide students with a benefit (in terms of attainment) in other, quantitative, Key Stage 5 subjects (e.g., A Level Psychology, BTEC Engineering)?

Data and methods

The main source of data for this research was the National Pupil Database (NPD) Key Stage 5 (KS5) extract for 2021/22. The NPD is administered by the Department for Education (DfE) and includes examination results for all students in schools and colleges in England. It also includes student and school background characteristics such as gender, ethnicity, prior attainment, and school type. We restricted the analysis to students who took at least one qualification equivalent in size to an A Level and who were aged 17 or 18 at the start of the academic year. We requested 2021/22 data, as this was the most recent available data. We acknowledge that in 2021/22 England was still coming out of a period in which exams were cancelled and school had been disrupted by the COVID-19 pandemic. However, fundamentally the 2021/22 academic year was more "normal" than the prior two academic years so provides a reasonable comparison to the analysis of data from pre-2020 years.

For research question 1, we analysed the background characteristics of CM students and compared this with the characteristics of non-CM students. The characteristics we looked at were prior attainment, gender, deprivation, ethnicity, first language, special educational needs (SEN), school type and school gender composition.

For prior attainment, we split the KS5 cohort of students into three equally sized groups ("High", "Medium", "Low") based on their average point score (APS) at Key Stage 4 (KS4). Average point score was calculated by assigning a point score to each achieved grade[3] and averaging this across all KS4 qualifications taken by the student.

3 E.g., for GCSEs the point score was the same as the grade (e.g., 9, 8, etc.). See https://www.gov.uk/government/publications/key-stage-4-qualifications-discount-codes-and-point-scores for details.

Student deprivation was measured by the Income Deprivation Affecting Children Index (IDACI), which indicates the proportion of children in the area a student lives in living in low-income families.[4] The KS5 cohort were split into three equally sized groups based on their IDACI score ("High", "Medium", "Low").

We used the ethnicity categories already recorded in the NPD to group students. These were Asian, Black, Chinese, Mixed, White, Other, and Unclassified. Chinese students were in a category of their own due to their tendency to achieve high grades compared to other Asian students. Students were also grouped by their first language (English or other).

For the students with SEN, we used the categories in the NPD. These were "SEN, no statement", and "SEN, with statement", with the second of these requiring the most support.[5]

For these last four student characteristics (IDACI score, ethnicity, language, and SEN), there was a large amount of missing data (around 50 per cent). This was because these variables were collected as part of the school census, using information provided by schools. However, independent schools and colleges were not required to provide this information, leading to large amounts of missing data from these school types. Therefore, any analysis involving these characteristics was carried out just for those students with available data.

For the analysis by school type, schools were grouped into six categories: comprehensive (including academies and secondary moderns), sixth form colleges, further education / tertiary colleges, independent schools, selective schools, and other schools.

Students were also classified by the gender composition of the school they attended. This was derived from the percentage of girls in each school. If this was greater than 95 per cent then the school was categorised as a girls' school, if it was less than 5 per cent it was categorised as a boys' school. Otherwise, it was categorised as a mixed gender school.

For research question 2, we present descriptive statistics on the qualifications and subjects most commonly combined with CM. For this analysis we considered both the most common A Level subjects and the most common non-A Level subjects.

For research question 3, we were interested in whether CM helped students' performance in other subjects with a quantitative element taken at the same time. For this analysis we removed students who took either AS or A Level Maths, as they would not be eligible to take CM. This meant we were directly comparing students taking CM with those not taking any maths in KS5.

4 For further information on IDACI calculation, including definitions of children, families, and income deprivation, see Smith et al. (2015).
5 A "statement" of special educational needs is a legal document which outlines the educational needs of the child and how they will be met by the local education authority.

We investigated performance in the eight A Level subjects with a quantitative element most commonly combined with CM. We also chose five subjects from the range of BTECs equivalent in size to one A Level, and five subjects from the range of BTECs equivalent in size to three A Levels. Again, these were all subjects with a quantitative element. This analysis consisted of a series of regression models.

Regression analysis

For each A Level or BTEC subject we investigated for research question 3, we fitted logistic regression models predicting the probability of students achieving a particular grade or higher. We chose two different grades for each subject. These grades were chosen to represent two different points across the grade distribution: firstly, a high achieving grade, only attained by a minority of students; and secondly, a grade somewhere in the middle of the distribution, which was achieved by a substantial majority of the students. For A Levels, the dependent variables were achieving at least a grade A and achieving at least a grade C. For BTECs equivalent in size to one A Level, the dependent variables were achieving grade D* and achieving at least a grade D.[6] For BTECs equivalent in size to three A Levels, the dependent variables were achieving at least a grade D*D*D and achieving at least a grade MMM.

In each model, we included a variable which indicated whether the student had taken CM or not. This was our main variable of interest. A statistically significant parameter estimate for this variable would indicate that taking CM had a significant effect on the probability of achieving a particular grade or higher.

We used multilevel regression models, as these accounted for the clustering of students within schools. For a more detailed description of multilevel logistic regressions see Goldstein (2011). The general form of the models was as follows:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \ldots + \beta_l x_{lij} + u_j$$

where p_{ij} is the probability of student i from school j achieving the relevant grade or higher, x_{1ij} to x_{lij} are the independent variables (including the indicator of taking CM), β_0 to β_l are the regression coefficients, and u_j is a random variable at school level.

For each regression model, other contextual variables which could have had an impact on the outcome variable were included as independent variables. These were student gender, prior attainment, deprivation, ethnic group, first language, special educational needs (SEN) status, student total qualification size, school type, school gender composition, and school mean KS5 attainment.[7]

Most of these variables were described in detail in the previous section of this

6 In BTECs, the grades (from high to low) are Distinction* (D*), Distinction (D), Merit (M), and Pass (P).

7 The base categories (or reference groups) used in the regression analyses for the categorical variables were: female; White; first language English; no SEN; comprehensive (including academies and secondary moderns); and mixed sex.

article. In addition, the student total qualification size variable indicated the total size of the KS5 qualifications taken by each student, measured in A Level equivalents. For example, a student taking three A Levels would have a value of 3. Other qualifications were already assigned a size in the NPD (e.g., BTECs were equivalent in size to either one, two or three A Levels).

For the school KS5 attainment measure (centre KS5 point score), we calculated the average KS5 point score among all students in each school. The KS5 point score for each student was available in the NPD data and (as with the KS4 point score) was calculated by assigning a point score to each achieved grade[8] and averaging this across all KS5 qualifications taken by the student.

A backwards stepwise approach was used to decide on which variables to include in the final models. This method involves starting with a model which includes all possible variables and then removing statistically non-significant variables one by one until only the statistically significant variables remain. Statistical significance was evaluated at the 5 per cent level.

To ensure confidentiality of the data, statistical disclosure controls have been applied to the results (tables and graphs). In particular, counts below 10 and percentages based on counts below 10 have either been suppressed or merged with other counts/percentages.

Results

Uptake of Core Maths by background characteristics

In the 2021/22 NPD data there were 11 522 students who took Core Maths (out of a cohort size of 442,963). Core Maths should mainly be taken by students who achieved a grade 4 or higher at GCSE Maths but did not go on to take A Level Maths. We checked whether this was the case by calculating the GCSE Mathematics grade distribution of CM students (where this data was available). We compared this with the grade distribution of those taking AS or A Level Mathematics and with those not taking any level 3 mathematics qualification. The results are shown in Table 1.

Table 1: GCSE Maths grade distribution by post-16 maths option (% of students achieving each grade)

		GCSE grade							
Level 3 maths	**N**	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**<3**
Core Maths	11 034	2.6	10.2	21.5	29.6	27.0	8.6	0.7	0.1
AS / A Level	76 508	33.8	33.5	23.5	7.7	1.3	0.2	<0.1	<0.1
No maths	318 321	1.3	5.1	11.6	17.2	26.4	22.9	9.6	5.9

This shows that over 99 per cent of CM students achieved a grade 4 or higher and most (78 per cent) achieved grades 5 to 7. These are the types of students the qualification is targeted at. Students going on to take AS or A Level Maths were much higher attaining, with over 90 per cent achieving grade 7 or higher.

8 For example, a grade A* at A Level is worth 60 points, A grade is worth 50 points, down to a grade E (10 points) and a grade U (0 points).

Table 2 summarises the background characteristics of CM students and how these compare with non-CM students (including non-CM students taking AS or A Level Maths). This shows some substantial differences between the two groups in their background characteristics. For more details on the comparison between CM and non-CM students, see Gill (2024).

Table 2: Comparison of background characteristics of CM and non-CM students

Background characteristic	Summarised comparison of CM and non-CM students
Gender	CM students were relatively evenly split between females (47.9%) and males (52.1%). This compares with 53.3% female and 46.7% male for non-CM students.
Prior attainment	CM students were most likely to be in the middle attainment group (46.3%), followed by the high attainment group (32.7%). This meant they were somewhat higher attaining on average than non-CM students (33.5% low attaining, 33.0% medium attaining, 33.6% high attaining).
Deprivation	CM students were slightly more likely to be in the low deprivation group (38.2%) than in the medium (32.3%) or high (29.5%) deprivation groups. This meant they experienced less deprivation on average than non-CM students (33.0% low deprivation, 33.5% medium deprivation, 33.5% high deprivation).
Ethnicity	CM students were more likely to be white (74.4%), and less likely to be Asian (11.5%) or Black (5.3%) than non-CM students (65.8%, 15.3%, and 7.7% respectively).
First language	CM students were more likely to be first language English speakers (85.6%) than non-CM students (81.0%).
SEN status	Students with SEN made up 6.3% of CM students. This was almost identical to the proportion among non-CM students (6.4%).
School type	CM students were more likely to attend comprehensives / academy schools (51.0%), or sixth form colleges (22.6%) and less likely to attend Further Education (FE) colleges (12.9%) or independent schools (2.4%) when compared to non-CM students (36.3%, 17.2%, 29.1%, and 8.5% respectively).
School gender composition	Students taking CM were slightly more likely to attend mixed schools (94.5%) and slightly less likely to attend boys' schools (1.6%) than non-CM students (94.0% and 2.0% respectively).

Qualifications and subjects taken by Core Maths students

Table 3 presents the qualifications (and combinations of qualifications) most likely to be taken alongside CM. It shows that the highest proportion of CM students (44.4 per cent) combined it with three A Levels. The next most common qualifications combined with CM were one BTEC only, followed by two A Levels and one BTEC, and three A Levels and EPQ.

Table 3: Types and numbers of qualifications most commonly combined with Core Maths

Combination	No. of students	Per cent of CM students
3 A Levels only	5115	44.4
1 BTEC only	883	7.7
2 A Levels / 1 BTEC	713	6.2
3 A Levels / 1 EPQ	572	5.0
2 A Levels / 1 VRQ[9]	439	3.8
2 A Levels only	333	2.9
2 BTECs only	282	2.5
1 A Level / 1 BTEC	267	2.3
2 A Levels / 1 Cambridge Technical	253	2.2
1 EPQ / 1 VRQ	168	1.5

Table 4 presents the most common A Level subjects combined with CM. Eight out of the top 10 had some quantitative elements, for which CM may be useful. The third column in the table shows the percentage of CM candidates who took the subject. For example, just over 30 per cent of CM candidates also took Psychology A Level. The final column in the table shows the percentage of students taking the A Level subject who also took CM. The highest percentages were for Geography (5.1 per cent) and Biology (5.0 per cent).

Table 4: A Level subjects most commonly combined with Core Maths (students can take more than one subject)

Subject	No. of CM students	Per cent of CM students	Per cent of students taking subject
Psychology	3464	30.1	4.6
Biology	3151	27.3	5.0
Chemistry	1891	16.4	3.6
Business Studies	1845	16.0	4.8
Geography	1756	15.2	5.1
Economics	1241	10.8	3.5
Sociology	1211	10.5	2.8
History	1135	9.9	2.7
Physics	635	5.5	1.8
English Literature	610	5.3	1.9

9 VRQ = Vocationally Related Qualification. These are mainly introductions to an area of work, but do not develop a recognised competence or lead directly to employment. Examples include Applied Diploma / Certificate in Criminology (WJEC), and Diploma / Certificate in Financial Studies (London Institute of Banking & Finance).

Table 5 shows the most common non-A Level subjects taken alongside CM. The EPQ was the most popular, with 11.6 per cent of CM students. This was followed by two BTECs (Applied Sciences, and Business).

Table 5: Non-A Level subjects most commonly combined with Core Maths (students can take more than one subject)

Qualification	Subject	No. of CM students	Per cent of CM students	Per cent of students taking subject
EPQ	n/a	1 342	11.6	2.7
BTEC	Applied Sciences	861	7.5	5.5
BTEC	Business	669	5.8	2.4
VRQ	Criminology	595	5.2	3.0
BTEC	Engineering	535	4.6	8.3
BTEC	Information Technology	371	3.2	4.5
BTEC	Health Studies	323	2.8	1.5
BTEC	Sports Studies	297	2.6	2.0
Cambridge Technical	Information Technology	260	2.3	5.0
VRQ	Financial Studies	229	2.0	3.3

A further analysis explored the most common combinations of subjects taken alongside CM. The most common combination was A Levels in Biology, Chemistry, and Psychology, taken by 453 students (3.9 per cent of CM students). The second and third most common combinations were both single BTECs worth three A Levels: Engineering, taken by 271 students (2.4 per cent); and Applied Sciences, taken by 256 students (2.2 per cent). Six out of the top 10 combinations were A Levels only or A Levels with EPQ. All of these combinations included A Level Biology, four included A Level Chemistry, and four included A Level Psychology.

We also looked at the most popular combinations in a different way, by calculating the subjects with the highest percentage of students also taking CM (Table 6). This was restricted to subjects with at least 100 entries. This may give an indication of which subjects and qualifications teachers and students believed would most benefit from CM being taken alongside.

Table 6: Subjects with highest percentage of students taking Core Maths (at least 100 entries)

Qualification	Subject	No. of CM students	Per cent of students taking subject
OCR Cambridge Tech Extended Diploma	Engineering	45	30.6
OCR Cambridge Tech Diploma	Engineering	79	25.2
VRQ	Religious Education	25	17.2
BTEC National Extended Diploma	Manufacturing Engineering	22	15.6
OCR Cambridge Tech Extended Cert	Engineering	74	13.8
BTEC Level 3 National Certificate	Applied Sciences	32	13.0
BTEC Certificate	Manufacturing Engineering	16	11.1
A Level	Environmental Science	125	10.7
BTEC National Foundation Diploma	Engineering	118	10.3

The highest percentage was for the OCR Cambridge Technical Extended Diploma in Engineering, with 30.6 per cent of the students taking the subject also taking CM. Six out of these nine qualifications were in an engineering-related subject. It is surprising that the subject with the third highest percentage was a VRQ in Religious Education, as this is not a subject with any quantitative element. However, the number of candidates taking this qualification was low (145), so we should not read too much into this.

Do Core Maths students perform better in subjects which have a quantitative element than similar students not taking Core Maths?

As described earlier, for this analysis we explored performance in the most common A Level and BTEC subjects taken alongside CM which were deemed to have a quantitative element.

For each subject, we ran two sets of regression models predicting the probability of achieving:

- at least grade A and at least grade C for A Level subjects
- grade D* and at least grade D for BTECs equivalent in size to one A Level
- at least grade D*D*D and at least grade MMM for BTECs equivalent in size to three A Levels.

Within each grade we also fitted multiple models. Firstly, we fitted a model including all variables (both at student and school level) which were statistically significant ("all variables" model). Secondly, a model was fitted which excluded the census variables (IDACI, ethnicity, language, and SEN) and retained statistically significant non-census variables. This was called the "no census variables" model. As noted in the data and methods section, the census variables have large amounts of missing data. Therefore, by fitting a model excluding these we were able to include many more students and get a sense of whether this affected the results.

The key regression results are presented in Tables 7 to 9. These show, for each subject in each qualification, the parameter estimates for the variable indicating whether CM was taken or not.

The results for A Levels (Table 7) show a positive effect of taking CM for all subjects and grades apart from sociology. However, there were only a few subjects for which the effect was significantly different from 0. In terms of the models with all variables in, there were significant positive effects for biology (grades A and C), chemistry (grade C), and business studies (grade A). All these instances were also significant in the models without the census variables (and mostly only changed in value by a small amount). There were also two instances (business studies grade C, and economics grade A) where there was no significant effect of CM in the models with census variables but with a significant positive effect in the models without census variables.

There was one instance of a significant negative effect of taking CM, for sociology grade A (although in the model without the census variables this was no longer significant). This finding is examined further in the discussion section.

Table 7: Parameter estimates for Core Maths variable (A Level subjects, standard errors in parentheses)

Subject	Grade predicted	Number of students		Core Maths parameter estimate	
		All variables model	No census variables model	All variables model	No census variables model
Psychology	At least grade A	42 174	66 209	0.034 (0.065)	0.103 (0.053)
	At least grade C			0.130 (0.072)	0.105 (0.059)
Biology	At least grade A	26 091	39 409	0.232 (0.073)*	0.235 (0.059)*
	At least grade C			0.180 (0.067)*	0.132 (0.055)*
Chemistry	At least grade A	14 122	21 735	0.096 (0.092)	0.124 (0.075)
	At least grade C			0.188 (0.083)*	0.145 (0.068)*
Business Studies	At least grade A	18 208	31 529	0.250 (0.088)*	0.199 (0.072)*
	At least grade C			0.184 (0.105)	0.247 (0.084)*
Geography	At least grade A	18 186	27 391	0.166 (0.087)	0.051 (0.075)
	At least grade C			0.057 (0.099)	0.068 (0.086)
Economics	At least grade A	11 060	18 487	0.105 (0.107)	0.175 (0.088)*
	At least grade C			0.204 (0.120)	0.175 (0.097)
Sociology	At least grade A	26 205	40 812	-0.249 (0.105)*	-0.150 (0.085)
	At least grade C			0.116 (0.120)	0.052 (0.100)
Physics	At least grade A	26 091	39 409	0.345 (0.222)	0.188 (0.201)
	At least grade C			0.253 (0.138)	0.118 (0.119)

In these logistic regressions, the parameter estimates are hard to interpret as they are the log of the odds of achieving the grade or higher. However, we can convert these into probabilities for "typical" students to illustrate the size of these effects. The typical students we chose were those in the base category for each of the categorical variables and with a value of each continuous variable equal to the mean. Figure 1 compares the probabilities (for CM and non-CM students) of achieving the relevant grade (or higher) for each subject and grade with a significant CM effect (using the results of the "all variables" models). It shows that the differences in probabilities were all very small, despite being statistically significant.

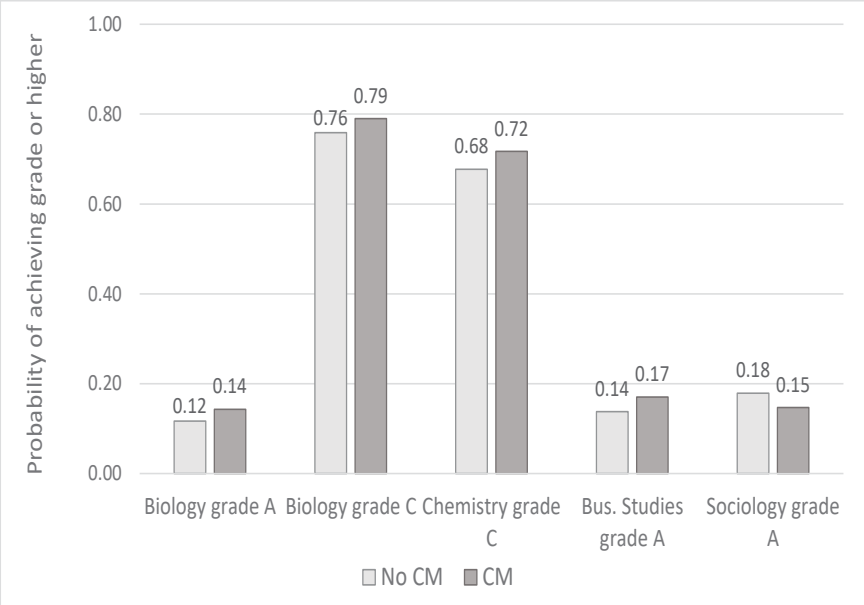

Figure 1: Probabilities of achieving a grade (or higher), for CM and non-CM students (A Levels; "all variables" models with significant CM effect)

The results for BTECs (equivalent in size to one A Level) are shown in Table 8. The "n/a" in the table means that for that particular combination of subject and grade none of the census variables had a significant effect, so there was no "all variables" model.

Table 8: Parameter estimates for Core Maths variable (BTEC subjects equivalent in size to one A Level, standard errors in parentheses)

Subject	Grade predicted	Number of students		Core Maths parameter estimate	
		All variables model	No census variables model	All variables model	No census variables model
Applied Sciences	Grade D*	3 373	4 577	0.492 (0.219)*	0.359 (0.176)*
	At least grade D			n/a	0.288 (0.144)*
Business	Grade D*	7 000	11 014	0.085 (0.214)	-0.080 (0.179)
	At least grade D			n/a	-0.005 (0.139)
Information Technology	Grade D*	3 314	5 211	n/a	-0.011 (0.203)
	At least grade D			0.165 (0.242)	0.233 (0.176)
Sport	Grade D*	3 883	5 453	n/a	-0.178 (0.230)
	At least grade D			0.060 (0.250)	-0.025 (0.216)
Health & Social Care	Grade D*	6 163	8 473	-0.058 (0.270)	-0.050 (0.211)
	At least grade D			0.183 (0.253)	-0.091 (0.209)

Only for one subject was there a significant effect of taking CM. This was applied sciences, which had significant positive effects for both grades. In terms of probabilities, "typical" CM students had a probability of achieving a grade D* in applied sciences of 0.10 compared with 0.06 for non-CM students, and a probability of achieving a grade D of 0.53 compared to 0.46 for non-CM students.

Table 9 presents the results for the BTECs equivalent in size to three A Levels. In all subjects there were no significant effects of the census variables. Therefore, the result of only one model (the "no census variables" model) is presented for each subject grade combination.

Table 9: Parameter estimates for Core Maths variable (BTEC subjects equivalent in size to three A Levels, standard errors in parentheses)

Subject	Grade predicted	Number of students	Core Maths parameter estimate
Applied Sciences	At least grade D*D*D	5 299	0.343 (0.198)
	At least grade MMM		0.614 (0.259)*
Engineering	At least grade D*D*D	2 478	0.108 (0.276)
	At least grade MMM		0.314 (0.269)
Information Technology	At least grade D*D*D	2 323	1.216 (0.407)*
	At least grade MMM		0.084 (0.349)
Business	At least grade D*D*D	7 886	-0.046 (0.316)
	At least grade MMM		0.720 (0.475)
Health & Social Care	At least grade D*D*D	7 206	-0.120 (0.508)
	At least grade MMM		-0.488 (0.493)

There were two subjects for which CM had a significant (positive) effect on performance. In applied sciences, this was for grade MMM or higher; in information technology, this was for grade D*D*D or higher. In terms of probabilities, "typical" CM students had a probability of achieving a grade MMM in applied sciences of 0.95 compared with a probability of 0.92 for non-CM students, and a probability of achieving a grade D*D*D in information technology of 0.21, compared with 0.07 for non-CM students.

Discussion

The main aims of this research were to investigate the position of the Core Maths qualifications in the KS5 curriculum, including uptake among students with different background characteristics and the qualifications and subjects it was combined with, and to see whether students taking CM performed better in their A Level or BTEC subjects taken at the same time.

The results showed that most students taking Core Maths in 2021/22 were those it was aimed at, i.e., achieving a grade 4 or higher in GCSE Maths, but not progressing to AS or A Level in the subject. Over 99 per cent of CM students achieved a grade 4 or higher in their GCSE, with most (78 per cent) achieving grades 5 to 7. On average, CM students achieved lower grades than AS/A Level students, but higher grades than those not taking any level 3 maths qualifications.

In terms of the background characteristics of CM students, we found the following:

- 52 per cent were female – this is a much more even split than in A Level Maths, which was 63 per cent male in 2021/22 (Gill, 2024). This suggests that CM could help with closing the gender gap in post-16 maths.
- CM students were less deprived than average, with 38 per cent in the "low" deprivation group (as measured by the IDACI).
- They were more likely than non-CM students to be white, first language English speakers and less likely to be Black or Asian or to have another first language.
- They were more likely to attend comprehensive schools, or sixth form colleges and less likely to attend FE colleges or independent schools when compared to all other students.

It was not within the scope of the current research to investigate the reasons for lower uptake levels in specific groups of students. Further research could investigate the reasons why particular groups of students were less likely to take CM (e.g., non-white, non-English speakers, those attending independent schools) and if anything can be done to encourage uptake among these groups of students.

However, there may also be a geographical aspect to this. Homer et al. (2020) noted that provision of CM throughout England was patchy. Many of the background characteristics we investigated (e.g., ethnicity, language, deprivation) are geographically clustered, and it may be that the areas where schools were less likely to offer CM were those with higher proportions of Black, Asian, second language English, or more deprived students. i.e., the problem is with provision of CM, not uptake.

CM students were most likely to combine the qualification with three A Levels (44 per cent of CM students). The next most common combination was with one BTEC (usually equivalent in size to three A Levels). The most common subjects combined with CM mostly had some quantitative element, such as A Level Psychology, Biology, and Chemistry, and BTEC Applied Sciences, Business Studies, and Engineering. These results suggest that CM was being taken by many students to support them in these other subjects. This confirms previous case study findings from Homer et al. (2020), who reported that several schools they surveyed required or strongly encouraged students taking particular subjects (e.g., BTEC applied sciences, A Level Psychology) to also take CM. Many students in their research also reported that they chose (or were required) to take CM because it would support them taking A Levels with a quantitative element.

The subjects with the highest proportions of students also taking CM were mostly Cambridge Technicals and BTECs. Six out of the top nine of these were engineering-related subjects. This suggests that this is a subject area where students were being particularly encouraged to take CM. This is not surprising, as engineering is a subject with a significant amount of mathematical content. It may be that students taking engineering were generally required to also take a level 3 maths qualification, either AS/A Level (for higher attainers) or CM (for lower attainers).

Although this research has shown that Core Maths is often taken alongside A Level and BTEC subjects with a quantitative component, there is still plenty of potential for increase in uptake. For example, Gill (2024) found that for some A Level subjects with high entries (e.g., Sociology, Psychology, Business Studies), there were still large percentages of students not taking any maths at all at KS5 (between 78.6 per cent and 93.2 per cent).

The current research provided some evidence that students taking CM achieved better grades than those not taking CM in some subjects with a quantitative element taken at the same time. The results of our analyses showed four occasions when CM students had a significantly higher probability of achieving a particular grade or higher in specific A Level subjects. This was for Biology grades A and C, Chemistry grade C, and Business Studies grade A. However, in each case the size of the effect was small (an increased probability of between 0.02 and 0.04).

Additional significant positive effects were identified for Business Studies grade C and Economics grade A, but only in the "no census variables" model. There was no obvious reason why these showed a significant effect while there was no such effect for the "all variables" model for these subjects and grades. One possible explanation is that the reduced sample in the "all variables" model excluded many of the students who benefitted from taking CM.

There was one significant negative effect of taking CM, for A Level Sociology. This reduced the probability of achieving at least a grade A for CM students from 0.18 to 0.15. It is not clear why taking CM was associated with worse performance in this subject, but it may reflect the relatively low levels of mathematical content in sociology. However, the size of the effect was very small.

These findings were somewhat different from those from previous research into the impact of taking CM on performance in other subjects. Homer et al. (2020) found no significant positive effects across five A Level subjects (Psychology, Biology, Business Studies, Geography, and Chemistry). Their only significant effect was a small negative one for A Level Business Studies. There are a number of possible explanations for this difference which relate to the qualification running for several more years since the last research was published, for example: the increase in uptake of CM in recent years; teachers having more experience of teaching the qualification; and schools being better at deciding which students CM is likely to help. Furthermore, the outcome variable in the previous research (point score achieved in the A Level) was different from the one in our research and their statistical model included fewer variables.

We also found evidence of an effect of taking CM on BTEC performance. For example, for BTECs equivalent in size to one A Level there were two significant positive effects on performance (applied sciences at grade D* and at grade D or above). Similarly, for BTECs equivalent in size to three A Levels, there were two significant positive effects (applied sciences at grade MMM or above; information technology at grade D*D*D or above). Two of these effects were very small, but two were substantially larger than the significant A Level effects. For applied sciences (worth one A Level), taking CM increased the probability of achieving grade D or better from 0.46 to 0.53. For information technology (worth three A Levels), taking CM increased the probability of achieving grade D*D*D or better from 0.07 to 0.21.

Overall, the positive effects of taking CM were mostly very small, but it is worth noting that several of them were in science subjects, which may have more mathematical content than the social science subjects we investigated (e.g., Sociology, Geography). It is also important to note that while the subjects we investigated had a quantitative element, for most of these the amount of mathematical content was not substantial, so it is probably unrealistic to expect to find large effects. One possible area of further research would be to look at question papers for subjects with a quantitative element and identify items requiring mathematical knowledge or skills, and then investigate if students taking CM performed significantly better on these items than non-CM students.

It should be noted that we need to be somewhat cautious with the interpretation of the results. Although, in some instances, we found a significant association between taking CM and achievement in other subjects taken concurrently, this does not mean that there was a causal link. There may be other reasons why CM students performed better. For example, it may be that students taking CM were more motivated to do well academically than non-CM students and it was this that meant they did better in their other subjects, rather than taking CM per se.

While this research suggests that CM could be having a positive impact for learners who take it, the issue of relatively low uptake amongst target learners remains, with only 11 522 entries in 2021/22 (amongst the 442 963 completing KS5 in that year). This would appear to be lower than was hoped, given that the development of these qualifications was aimed at the 200 000 students who

achieved a grade C in Maths GCSE but did not go on to AS or A Level Maths (DfE, 2013). It is worth noting that in February 2024 the Education and Skills Funding Agency announced the "Core Maths premium", which is additional funding for CM students to support the planned introduction of the Advanced British Standard (ESFA, 2024). It will be interesting to see whether this has any impact on uptake levels. There is certainly scope for greater numbers of students to take advantage of the potential benefits of studying the qualification, particularly amongst groups of students where there is currently lower uptake.

References

AMSP. (no date). *Level 3 maths update 2023-24*. Advanced Mathematics Support Programme.

DfE. (2013). *Introduction of 16 to 18 core maths qualifications. Policy statement*. Department for Education.

DfE. (2017). *Key stage 4 shadow measures*. Department for Education.

ESFA. (2024). *Guidance 16 to 19 funding: Core maths premium.* Education & Skills Funding Agency.

Gill, T. (2024). *Core Maths qualifications: How they fit in post-16 programmes of study and their impact on other subjects with a quantitative element*. Cambridge University Press & Assessment.

Goldstein, H. (2011). *Multilevel Statistical Models (4th edition*). John Wiley & Sons.

Homer, M., Mathieson, R., Tasara, I., & Banner, M. L. (2020). *The early take-up of Core Maths: Successes and challenges*. University of Leeds.

Royal Society. (2023, December 13). *Why Core Maths?* https://royalsociety.org/topics-policy/projects/why-core-maths/

Smith, T., Noble, M., Noble, S., Wright, G., McLennan, D., & Plunkett, E. (2015). *The English Indices of Deprivation 2015 Technical report*. Department for Communities & Local Government.

Does typing or handwriting exam responses make any difference? Evidence from the literature

Santi Lestari (Research Division)

Introduction

Computer-based tests have become widespread in many assessment contexts, including language assessments and university admissions tests. Many general qualifications exams, however, remain in a paper-based mode, often requiring students to handwrite long answers, such as essays, under time constraints. Insufficient and unequal digital provision across schools is often identified as a major barrier to a full adoption of computer-based exams for general qualifications in many jurisdictions, including in England (Coombe et al., 2020). One feasible approach to overcoming this barrier is a gradual adoption, which involves offering both modes of exam administration in parallel (i.e., paper-based and computer-based) (Arce-Ferrer & Bulut, 2018; Coombe et al., 2020). This approach, however, presents risks of mode effects (Coombe et al., 2020). Mode effects occur when there are unavoidable differences between paper-based and computer-based exams that are intended to be equivalent. This can mean the exams measure slightly different constructs and the resulting scores may not be directly equivalent. When an exam is offered in both paper-based and computer-based modes, and results from both are treated as equivalent, and therefore interchangeable, the comparability between modes needs to be ascertained. This includes investigating potential response mode effects for extended writing questions, or, in other words, examining whether the mode in which students respond to the questions (i.e., by handwriting or typing on the computer) introduces systematic differences. We conducted a literature review on writing response mode effects, and this article summarises the key findings.

Methods

To identify the relevant studies, we searched major databases in education, psychology and linguistics, including Education Resources Information Center (ERIC) and Linguistics and Language Behavior Abstracts (LLBA). Keywords used in the searches included words related to i) writing mode such as “typed”, “typing”, “word-processed”, “handwritten” and “handwriting”, and ii) assessment such as “exam”, “examination”, “test” and “exam script”. We also checked the reference lists of the selected studies to find additional studies.

The criteria for inclusion in the review were that studies had to: a) be published in English; b) compare the two writing modes (i.e., handwriting and typing/word processing) in an assessment context; c) involve an assessment that required an extended writing response; and d) involve empirical data (i.e., using students' writing performance data from either an operational exam administration and/or an experimental setting). We decided to include various publication types (i.e., peer-reviewed journal articles, conference papers, doctoral theses/dissertations and institutional reports). This is because research investigating mode effects, especially for high-stakes assessments, is often conducted by awarding organisations and published only as an institutional report. We read the selected studies to identify the research context, focus and key findings.

Findings

Overview of the studies included

A total of 47 studies, published between 1990 and 2021, were included in the review (Figure 1). These studies varied in terms of context and focus. Figure 2 summarises the number of studies by research context. Almost half of the studies (22 out of 47) were conducted in language assessment contexts, almost exclusively in English as a second or foreign language (ESL/EFL) contexts (e.g., Brunfaut et al., 2018; Chan et al., 2018; Lessien, 2013; Manalo & Wolfe, 2000a), with only one study investigating mode effects in another language, namely Mandarin Chinese as a foreign language (Zhu et al., 2016). It is not surprising that language assessment is the dominant context given that writing as part of language proficiency is commonly tested in direct language assessments.[1] Some of the ESL/EFL assessments are also high-stakes in nature because important, often life-changing, decisions are made based on the test scores, giving more reason to investigate potential mode effects.

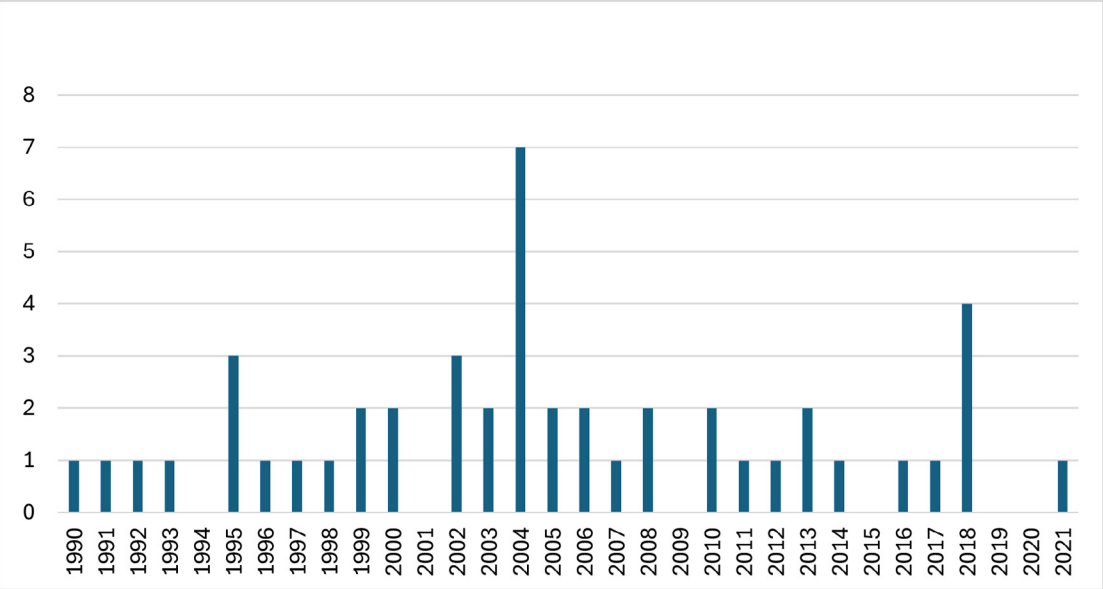

Figure 1: Number of studies across the years (n=47)

1 As opposed to indirect language assessments which measure writing proficiency through means other than directly requiring candidates to write, e.g., error recognition.

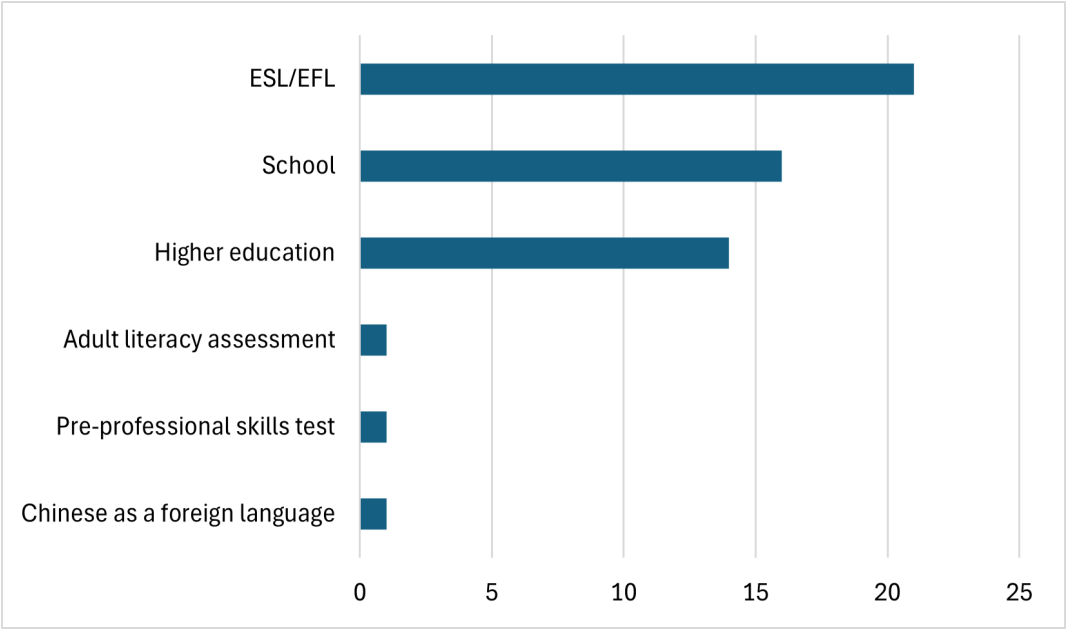

Figure 2: Number of studies by research context (n=47, multiple contexts possible)

In terms of level of education, 16 studies were conducted in school contexts, the majority of which were in the US (e.g., Burke & Cizek, 2006; Hollenbeck et al., 1999; Russell & Tao, 2004b; Wolfe et al., 1995; Wolfe et al., 1996). There are also school-based studies investigating mode effects in other jurisdictions including the UK (Charman, 2014; Connelly et al., 2007), Australia (MacCann et al., 2002) and Hong Kong (Lam & Pennington, 1995). Fourteen studies were conducted in higher education contexts. Such studies might focus on ESL/EFL (e.g., Jin & Yan, 2017; Kim et al., 2018), a non-language subject, such as theology (e.g., Mogey & Hartley, 2013) or admissions tests (e.g., Bridgeman & Cooper, 1998). Two studies do not fit into these education levels: Chen et al. (2011) studied mode effects of adult literacy assessment in the US, called the National Assessment of Adult Literacy (NAAL), and Yu et al. (2004) examined mode effects of the essay writing component of the Praxis Pre-Professional Skills Test, a battery test assessing basic academic skills of pre-service teachers.

Studies also varied in terms of the focus of their investigation (Figure 3). The primary focus of most studies was on the comparability of students' performance across the two modes of writing. Most studies operationalised performance as scores (e.g., Lam & Pennington, 1995; Yu & Iwashita, 2021), but some also examined the comparability of the characteristics of the texts produced (e.g., Barkaoui & Knouzi, 2018; Chambers, 2008; Charman, 2014; Jin & Yan, 2017) and a few investigated the comparability of students' composing processes across the two modes (Chan et al., 2018; Jin & Yan, 2017; Lee, 2002; Wolfe et al., 1993).

Researchers examining the comparability of scores across the two writing modes also often gathered students' contextual information, including demographic data such as gender, ethnicity and socio-economic background (e.g., Bridgeman & Cooper, 1998; Chen et al., 2011), language proficiency level (e.g., Lessien, 2013;

Manalo & Wolfe, 2000a) and information on students' computer familiarity[2] and/or perceptions of the composition mode (e.g., Barkaoui & Knouzi, 2018; Jin & Yan, 2017; Whithaus et al., 2008; Wolfe et al., 1996). Contextual information is useful to allow more fine-grained analyses of writing mode effects across sub-groups of a candidate population or to explain the presence of mode effects, if any.

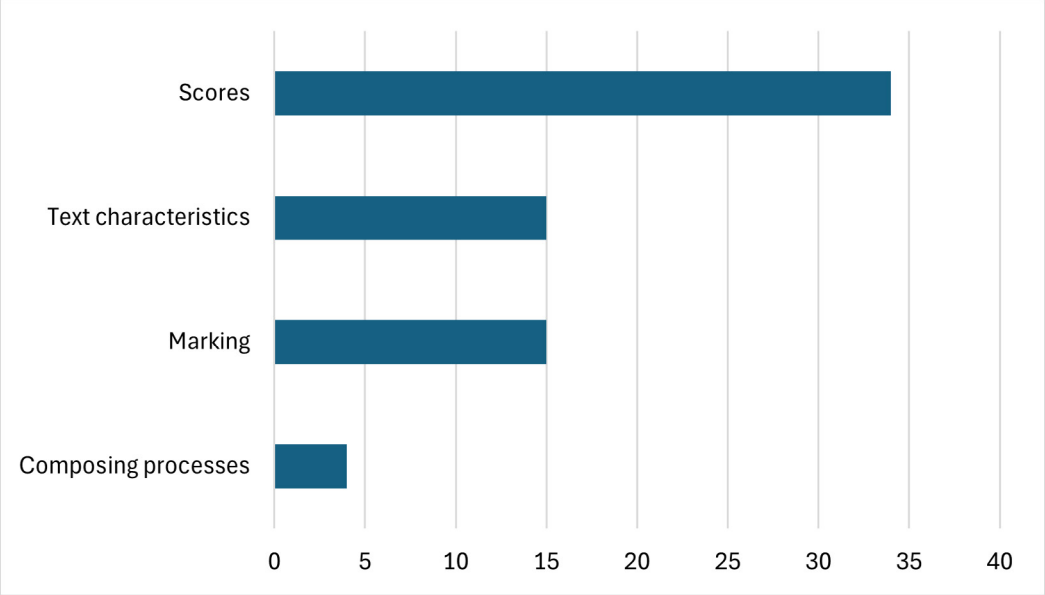

Figure 3: Number of studies by research focus (n=47, multiple focuses possible)

Another research focus was the effect of presentation mode on essay marking. More precisely, the research typically looked into whether the mode of script presentation to markers (i.e., handwritten or typed/word-processed) had differential effects on marking quality including marker bias (e.g., Arnold et al., 1990; Brown, 2003; Chen et al., 2011), marking processes (Wolfe et al., 1993), and other measures of marking quality such as inter-rater agreement and/or reliability (Lee, 2004; Manalo & Wolfe, 2000b).

The following sections present key findings under each research focus.

Comparability of scores

As the research methods used to investigate the comparability of scores vary considerably, it is important to be cautious in drawing conclusions from different research findings. Arce-Ferrer and Bulut (2018) examined four commonly used data collection designs[3] in mode effects studies and concluded that the single-group design with counterbalancing and random-groups design were the superior data collection designs at detecting mode effects at the test level (i.e., score distributions). Furthermore, if a score comparison is made at the group level, rather than the individual level, the score comparability conclusion may also hold true at the group level only. It is typically the case with the studies included in this

2 The term "computer familiarity" is used in the current article to include typing skills, word processing skills, experience or frequency of using a computer and level of comfort or confidence in using a computer.
3 Single-group design with counterbalancing, single-group design without counterbalancing, random-groups design, and anchor-test-nonequivalent-groups design.

review that the score comparability analysis was conducted at the group level rather than individual level, for example, by comparing the mean scores of each group.

Typed essays scored higher than handwritten essays

Some studies found that students performed better when they typed or word-processed their essay than when they handwrote it (Lam & Pennington, 1995; Lessien, 2013; Russell & Haney, 1997; Russell & Plati, 2002; Zhu et al., 2016). Russell and Haney (1997) and Lessien (2013) also found that the writing mode effect was highly significant, favouring typed essays, and this was particularly the case for students with high proficiency in English (Lessien, 2013). Findings from Zhu et al. (2016) were particularly interesting as this study investigated writing mode effects in Mandarin Chinese as a foreign language. Most of the students in the study also reported that they preferred word-processing their essay to handwriting it, as they felt word-processed essays appeared more professional.

Typed essays scored lower than handwritten essays

Other studies found that students performed better when they handwrote their essay than when they typed it (Breland et al., 2004; Bridgeman & Cooper, 1998; Chen et al., 2011; Connelly et al., 2007; Green & Maycock, 2004; Manalo & Wolfe, 2000a; McGuire, 1995; Yu et al., 2004). Manalo and Wolfe (2000a) found that when language proficiency was controlled for, the handwritten essays were scored approximately one-third of a standard deviation higher than the typed essays. Researching writing mode effects among primary school students aged 4 to 11 years old, Connelly et al. (2007) found that the quality of the handwritten scripts was better than that of the typed scripts. A differential effect of writing mode was also observed in Chen et al. (2011), whereby the computer-based mode disadvantaged unemployed candidates even more than employed candidates. Bridgeman and Cooper (1998), using the Graduate Management Admission Test (GMAT) essay task, observed that the score difference favouring handwriting mode did not interact with candidates' gender, ethnicity or English as a second language group classification.

No meaningful score difference between typed and handwritten essays

Additional studies found that generally there were no (meaningful) writing mode effects on students' performance (Barkaoui & Knouzi, 2018; Brunfaut et al., 2018; Chan et al., 2018; Charman, 2014; Horkay et al., 2006; Lee, 2002; Lovett et al., 2010; MacCann et al., 2002; Mogey et al., 2010; Yu & Iwashita, 2021). For instance, Chan et al. (2018), investigating the comparability of paper-based and computer-based delivery of the IELTS Writing test, found that scores across both modes were generally comparable although candidates scored better in the Lexical Resources criterion when they handwrote their essay. Chan et al. (2018) theorise that different writing modes might elicit certain aspects of writing, in this case lexical resources, slightly differently. Furthermore, they also observed that some aspects of computer familiarity significantly predicted performance in computer-based writing assessment, confirming findings from an earlier study by Horkay et al. (2006).

Mode effects and contextual variables

Writing mode effects are not always straightforward and can be influenced by the students' contextual factors and methods used in scoring the writing. Students' computer familiarity and typing speed were found to interact with writing mode (Russell, 1999; Wolfe et al., 1995). Students with greater familiarity with word processing software tended to perform equally well either typing or handwriting their essay, whereas students with less word processing experience tended to perform better and write more when handwriting their essay (Wolfe et al., 1995). Students' language proficiency is another factor that may influence writing mode effects. Students with weaker English language ability tended to perform better on handwritten essays, while those with better English performed comparably on both writing modes (Wolfe & Manalo, 2004). A similar finding was also observed by Brunfaut et al. (2018) in that the student group taking the lowest level of the English proficiency tests found a writing task easier in the handwriting mode than in the typing mode. Scoring method (i.e., holistic versus analytic) was also found to influence the scores of writing produced under the two writing modes. When holistic rating was used, no significant mean score difference was observed across the two modes; however, word-processed essays received significantly higher scores when analytic scoring was used (Lee, 2004).

Comparability of marking

The focus of marking comparability is on the effect of essay presentation mode on the marker (i.e., whether markers give different scores to the handwritten and typed versions of the same essays). Marker bias (i.e., whether markers give systematically higher scores on one presentation mode over another) was the primary focus of most studies examining comparability of marking across the two presentation modes. A few studies, however, also focused on the comparability of inter-rater agreement and reliability across the two modes. Some studies examined markers' perceptions of scoring essays in the two modes.

Marker bias

Handwritten essays were generally found to receive higher scores than the typed or word-processed versions of the same essays (Arnold et al., 1990; Brown, 2003; MacCann et al., 2002; Powers et al., 1994; Russell & Tao, 2004a; Shaw, 2003; Sweedler-Brown, 1991). The magnitude of the marker bias sometimes varied across different levels of performance. For example, Sweedler-Brown (1991) found that marking bias was more prominent for higher level performance; there was a significant difference in scores between modes for essays that received higher scores in the original handwritten format, but not for essays that received lower scores in the original handwritten format. Brown (2003) also found that the bias effect was moderated by the legibility of the handwriting, in that the score difference was higher for essays with poor legibility. This suggests that students with poor handwriting were, surprisingly, somewhat advantaged.

Chen et al. (2011), conversely, found no statistically or practically significant difference in the scores awarded to the typed and handwritten versions of essays. Similarly, Green and Maycock (2004) found that presentation mode effect was only negligible and of no practical importance.

Several potential explanations were identified for the common finding of bias against typed essays. Markers tended to have a higher expectation of word-processed essays (Arnold et al., 1990; Russell & Tao, 2004a). Word-processed essays were also often perceived to be shorter than handwritten essays although they were exactly of the same length (Arnold et al., 1990; Powers et al., 1994). Altering formatting style such as space and font size to make the word-processed essays appear to have a similar length to the handwritten version was found to reduce the size of presentation mode effect in Powers et al. (1994) but not in Russell and Tao (2004a).

Although word-processed essays were found to be easier to read, surface errors such as spelling and punctuation errors tended to appear more prominent and therefore more recognisable (Arnold et al., 1990; Russell & Tao, 2004a; Shaw, 2003; Wolfe et al., 1993). Handwriting, especially poor handwriting, could also mask such errors (Powers et al., 1994), which might explain Brown's (2003) finding above. Some markers in Russell and Tao (2004a) also reported that they could see students' effort more in handwritten essays, echoing findings from Powers et al. (1994) suggesting that traces of revisions in handwritten essays, such as strikethroughs, seemed to be valued by markers (who were usually also teachers). These factors may explain the bias against word-processed essays.

Marking reliability

Markers were generally found to have stronger agreement when scoring essays in the word-processed format than in the handwritten format. For example, Lee (2004) found that markers reached higher percentages of exact agreement when marking word-processed (76.1 per cent) and transcribed essays (78.6 per cent) than when marking handwritten essays (64.3 per cent). Furthermore, using other measures of inter-rater agreement and reliability (i.e., Pearson product moment correlation and Cohen's kappa), Manalo and Wolfe (2000b) and Wolfe and Manalo (2005) found that it was easier for markers to agree on scores for the word-processed essays than for the handwritten ones. Markers in Shaw (2003) reported that word-processed essays had a more similar general appearance and that both strong and weak essays were easier to read, potentially contributing to the increased objectivity.

Differences in scoring processes

The analysis of think-aloud protocol data in Wolfe et al. (1993) revealed differences in the processes involved in marking handwritten and word-processed scripts. When reading the handwritten essays, markers read less at a time and paused more often to make evaluative comments about the essay. In contrast, when reading the word-processed essays, they paused less frequently and saved most of the comments until after finishing reading the entire essay. Commentary on the word-processed essays tended to focus on the development of the essay, while comments on the handwritten essays focused more on essay organisation and authorial voice.

Comparability of text characteristics

Text length, typically measured in word and/or sentence count, is the most common measure of text characteristics explored in the studies that were

reviewed. Students tended to write longer texts when using a computer than when writing by hand (e.g., Barkaoui & Knouzi, 2018; Jin & Yan, 2017; Kim et al., 2018; Lee, 2002; Lovett et al., 2010; Mogey et al., 2010; Russell & Haney, 1997). However, this difference was not always statistically significant. The use of the keyboard could potentially explain the increased fluency in computer-based writing tests (Kim et al., 2018). Some studies found that text length also varied more considerably in word-processed essays than in handwritten ones (e.g., Chen et al., 2011; Endres, 2012).

In terms of language complexity, word-processed essays were found to have higher lexical variation (Barkaoui & Knouzi, 2018; Chambers, 2008; Charman, 2014), and more sophisticated vocabulary and varied syntactic structures (Barkaoui & Knouzi, 2018). Kim et al. (2018) also found similar patterns of results but commented that the differences were unlikely to be meaningful as average differences were relatively small and there was considerable overlap in values between the two modes.

Errors, usually mechanical errors such as punctuation and capitalisation, were also an area of investigation under text characteristics. It was generally found that there were no major differences in terms of the frequency of errors in handwritten and word-processed essays (e.g., Chambers, 2008; Endres, 2012; Wolfe et al., 1996). However, the nature of errors might differ. For example, Endres (2012) found that spelling errors in computer-based English writing tests were mainly typographical errors, which were potentially caused by typing errors, whereas spelling errors in the equivalent paper-based tests tended to be more developmental errors, potentially resulting from first language interference. Jin and Yan (2017), however, found that students made significantly fewer errors when they typed their essays than when they handwrote them, even though editing tools, such as grammar- and spell-checkers, were disabled.

Other features of text characteristics examined in previous studies include tone and readability. Whithaus et al. (2008) found that informal tone was perceived to be less present in typed essays than in handwritten ones. Using various readability indices including Flesch Reading Ease scores and Fog index, Mogey and Hartley (2013) found that the typed essays were generally more readable than the handwritten ones.

Most studies examining the comparability of text characteristics, however, did not consider students' level of computer familiarity. Including this aspect in their study, Wolfe et al. (1996) found that using a word processor did not impact the writing quality of students with medium and high levels of computer familiarity, but it harshly impacted those with lower levels of computer familiarity. On text length, specifically, students with medium and high levels of computer familiarity wrote longer word-processed essays than handwritten essays. In contrast, students with low familiarity wrote over 100 words fewer on average on a word processor than on paper. Furthermore, students with a medium or high level of computer familiarity tended to write a higher number of simple sentences when handwriting their essays compared to when typing them. Conversely, those with a low level

of computer familiarity tended to write more simple sentences when typing compared to when handwriting their essays.

In summary, differences in terms of text characteristics were observed between typed and handwritten essays. These differences, however, were not always statistically significant and/or of practical importance, and, furthermore, were not necessarily reflected in scores (Barkaoui & Knouzi, 2018). It should also be noted that writing modes might have differential effects on students with different levels of computer familiarity, as Wolfe et al. (1996) observed.

Comparability of composing processes

Composing processes refer to the activities that students engage in when answering an extended writing question. Chan et al. (2018) and Jin & Yan (2017) found that both writing modes elicited similar composing processes. However, a few differences were observed. In Jin and Yan's (2017) study, students with low and moderate levels of computer familiarity admitted that they planned better when handwriting their essay in the paper-based mode. One candidate explained that as they were required to handwrite their essay using a pen in the paper-based mode, they were more inclined to plan more carefully before writing to avoid making many corrections during writing, which would affect the essay presentation. In contrast, typing their essay on the computer allowed them to review and edit their essay more flexibly and therefore they were less inclined to plan more carefully before writing (Jin & Yan, 2017). Similarly, Chan et al. (2018) found some minor differences especially in planning, generating texts and monitoring and revising, although these differences in composing processes might not necessarily be reflected in scores. In terms of revising, some students in the study reported that when handwriting their essay in the paper-based mode, they tended to focus more on word level revisions, but when typing their essay in the computer-based mode, they tended to revise at the clause and sentence levels. Again, these differences were likely to be due to the flexibility afforded by the computer-based mode.

Discussion and conclusion

The question of whether typing or handwriting answers to extended writing questions in exams makes a difference has been widely investigated although the context and focus on which research has been conducted varied. In terms of context, more studies have been carried out in the context of English as a second or foreign language assessment, including proficiency and placement tests in higher education settings. Studies in the context of school education have been conducted in the US more than in any other jurisdiction, although this could be due to publication bias as we selected only articles and reports published in the English language. In terms of research focus, four aspects of comparability have been investigated: scores, marking, text characteristics and composing processes.

For **comparability of scores**, we could see that more studies, particularly the recent ones (which often used more robust methods involving the single-group design with counterbalancing and controlling for contextual factors), tended to find that scores across the two writing modes were comparable, at least at

the group level. However, there were also non-trivial numbers of studies that found a mode effect in one direction or the other. In a few studies, two contextual factors have been found to interact with mode effects: English proficiency and computer familiarity. Students with weaker English language ability tended to perform better on handwritten essays, while those with better English performed comparably on both modes. This particularly concerns writing mode effects in the context of ESL/EFL assessments. One implication is that when designing tests targeted specifically at students with low language proficiency, test designers may need to carefully consider whether to require students to type their essay as typing may underestimate the measurement of their writing ability.

Students with greater familiarity with word processing software tended to perform equally well either typing or handwriting their essay, whereas students with less experience with word processing tended to perform better and write more when handwriting their essay. It is therefore important to ensure that students have a sufficient level of computer familiarity, especially typing and word processing skills, to perform the assessment tasks. When it is known that a candidate pool varies considerably in their level of computer familiarity, it is recommended for test developers to offer both options of writing mode. However, as computer literacy is considered an indispensable aspect of academic literacy in the 2Ist century, some may argue that computer literacy should be considered an important element of the construct measured both in language assessment and in the assessment of other subjects (see e.g., Jin & Yan, 2017).

In terms of **comparability of marking**, handwritten essays generally appeared to receive higher scores than the word-processed version of the same essays. Reasons for this include markers having a higher expectation of word-processed essays and that word-processed essays were often perceived to be shorter than the handwritten version. As word-processed essays are easier to read, surface and mechanical errors such as spelling and punctuation become more recognisable to markers. On the other hand, handwriting, especially with low legibility, could mask such errors. Markers (who are usually teachers) also seemed to appreciate traces of corrections in handwritten essays such as strikethroughs, further contributing to bias against typed essays.

One possible measure to reduce such bias is through training. If exams are offered in both writing modes, it might be possible to train markers to ignore differences pertaining to each mode. However, there remain very limited studies on the effectiveness of training in reducing presentation mode effects on marker bias.

One important caveat to keep in mind regarding the literature on mode bias in marking, is that most of the relevant studies are at least 20 years old and took place before on-screen marking of scanned paper exam scripts became common practice. Given some of the possible contributors to bias relate to handwriting and legibility, which would be visible in scans of handwritten essays, there is still potential for there to be bias in current marking. On the other hand, markers' expectations of students' word-processed essays might have changed over time. Further evidence on whether bias against typed essays is present in current

marking, including when both handwritten and typed essays are marked on screen, would be valuable.

For **comparability of text characteristics**, the most frequent characteristic compared was text length. Word-processed essays tended to be longer than handwritten essays. However, the length of word-processed essays also appeared to vary more than that of handwritten essays. As computer familiarity could affect the length of essays produced, caution must be exercised to mitigate any risk of markers being biased by essay length. Although essay length has often been found to strongly correlate with scores and/or to be a strong predictor of scores (see Jeon & Strube, 2021; Kobrin et al., 2007), it is an irrelevant construct to writing. If Artificial Intelligence (e.g., an automated essay scoring system) is used for marking, it is crucial to ensure that the system does not rely on essay length in generating scores (see Jeon & Strube, 2021; Madnani & Cahill, 2018; Perelman, 2014). Using an automated scoring system that relies on construct-irrelevant features, including essay length, could threaten the interpretation of scores generated by the system (Bejar, 2017). Other differences in text characteristics such as language complexity and frequency and type of errors were also observed, but they were usually of little practical significance and may not necessarily translate to score differences.

There is a dearth of research examining **the comparability of composing processes** under the two writing modes. The few existing studies indicated that both modes elicit comparable processes with some minor differences. Comparable composing processes imply that both writing modes activate similar cognitive processes from students while they are engaged in task completion. Establishing cognitive equivalence between modes of composition becomes crucial when both modes are made available and schools may choose a composition mode on which their students are going to take the test.

In conclusion, potential mode effects due to writing mode can generally be considered a mature field of inquiry, evidenced by the number of empirical studies included in this review. Variability in research contexts, focuses and methods also further evidences the maturity of the research area. Such variability partly explains the differences in findings presented in this article. It should also be noted that some studies included in this review were conducted quite a while ago. Therefore, the generalisability and applicability of the findings should be considered carefully, given that both students and markers are likely to have increased familiarity and comfort with using a computer. An important aspect of writing mode effects in exams that remains little explored is the congruence between mode of learning and mode of testing and the extent to which this could influence mode effects.

Acknowledgement

The author would like to thank Camilo Ramos for earlier discussions of the literature in this area.

References

Arce-Ferrer, A. J., & Bulut, O. (2018). Effects of data-collection designs in the comparison of computer-based and paper-based tests. *The Journal of Experimental Education, 87*(4), 661–679.

Arnold, V., Legas, J., Obler, S., Pacheco, M. A., Russell, C., & Umbdenstock, L. (1990). *Do students get higher scores on their word-processed papers? A study of bias in scoring hand-written vs. word-processed papers*. Rio Hondo College.

Barkaoui, K., & Knouzi, I. (2018). The effects of writing mode and computer ability on L2 test-takers' essay characteristics and scores. *Assessing Writing*, *36*, 19–31.

Bejar, I. I. (2017). Threats to score meaning in automated scoring. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments* (pp. 75–84). Routledge.

Breland, H., Lee, Y. W., & Muraki, E. (2004). *Comparability of TOEFL CBT writing prompts: Response mode analyses* (Research Report, Issue RR-75). Educational Testing Service.

Bridgeman, B., & Cooper, P. (1998, April 13–17). *Comparability of scores on word-processed and handwritten essays on the Graduate Management Admissions Test* [Paper presentation]. The Annual Meeting of the American Educational Research Association, San Diego, CA, United States.

Brown, A. (2003). *Legibility and the rating of second language writing: An investigation of the rating of handwritten and word-processed IELTS Task Two essays* (IELTS Research Reports, Issue 4). IDP: IELTS Australia.

Brunfaut, T., Harding, L., & Batty, A. O. (2018). Going online: The effect of mode of delivery on performances and perceptions on an English L2 writing test suite. *Assessing Writing*, *36*, 3–18.

Burke, J. N., & Cizek, G. J. (2006). Effects of composition mode and self-perceived computer skills on essay scores of sixth graders. *Assessing Writing*, *11*(3), 148–166.

Chambers, L. (2008). Computer-based and paper-based writing assessment: a comparative text analysis. *Research Notes*, *34*, 9–15.

Chan, S., Bax, S., & Weir, C. (2018). Researching the comparability of paper-based and computer-based delivery in a high-stakes writing test. *Assessing Writing*, *36*, 32–48.

Charman, M. (2014). Linguistic analysis of extended examination answers: Differences between on-screen and paper-based, high- and low-scoring answers. *British Journal of Educational Technology*, *45*(5), 834–843.

Chen, J., White, S., McCloskey, M., Soroui, J., & Chun, Y. (2011). Effects of computer versus paper administration of an adult functional writing assessment. *Assessing Writing*, *16*(1), 49–71.

Connelly, V., Gee, D., & Walsh, E. (2007). A comparison of keyboarded and handwritten compositions and the relationship with transcription speed. *British Journal of Educational Psychology*, *77*(2), 479–492.

Coombe, G., Lester, A., & Moores, L. (2020). *Online and on-screen assessment in high-stakes, sessional qualifications: A review of the barriers to greater adoption and how these might be overcome*. (Ofqual/20/6723/1).

Endres, H. (2012). A comparability study of computer-based and paper-based Writing tests. *Research Notes*, *49*, 26–33.

Green, T., & Maycock, L. (2004). Computer-based IELTS and paper-based versions of IELTS. *Research Notes*, *18*, 3–6.

Hollenbeck, K., Tindal, G., & Almond, P. (1999). Reliability and decision consistency: An analysis of writing mode at two times on a statewide test. *Educational Assessment*, *6*(1), 23–40.

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment*, *5*(2), n2.

Jeon, S., & Strube, M. (2021). Countering the influence of essay length in neural essay scoring. The second workshop on simple and efficient natural language processing. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, 32–38, Virtual. Association for Computational Linguistics.

Jin, Y., & Yan, M. (2017). Computer literacy and the construct validity of a high-stakes computer-based writing assessment. *Language Assessment Quarterly*, *14*(2), 101–119.

Kim, H. R., Bowles, M., Yan, X., & Chung, S. J. (2018). Examining the comparability between paper- and computer-based versions of an integrated writing placement test. *Assessing Writing*, *36*, 49–62.

Kobrin, J. L., Deng, H., & Shaw, E. J. (2007). Does quantity equal quality? The relationship between length of response and scores on the SAT essay. *Journal of Applied Testing Technology, 8*(1), 1–15.

Lam, F., & Pennington, M. C. (1995). The computer vs. the pen: A comparative study of word processing in a Hong Kong secondary classroom. *Computer Assisted Language Learning*, *8*(1), 75–92.

Lee, H. (2004). A comparative study of ESL writers' performance in a paper-based and a computer-delivered writing test. *Assessing Writing*, *9*(1), 4–26.

Lee, Y.-J. (2002). A comparison of composing processes and written products in timed-essay tests across paper-and-pencil and computer modes. *Assessing Writing*, *8*(2), 135–157.

Lessien, E. (2013). *The effects of typed versus handwritten essays on students' scores on proficiency tests* [Unpublished doctoral dissertation, Michigan State University].

Lovett, B. J., Lewandowski, L. J., Berger, C., & Gathje, R. A. (2010). Effects of response mode and time allotment on college students' writing. *Journal of College Reading and Learning*, *40*(2), 64–79.

MacCann, R., Eastment, B., & Pickering, S. (2002). Responding to free response examination questions: Computer versus pen and paper. *British Journal of Educational Technology*, *33*(2), 173–188.

Madnani, N. & Cahill, A. (2018). Automated scoring: Beyond natural language processing. *Proceedings of the 27th International Conference on Computational Linguistics,* 1099–1109.

Manalo, J. R., & Wolfe, E. W. (2000a, April 24–28). *A comparison of word-processed and handwritten essays written for the Test of English as a Foreign Language* [Paper presentation]. The Annual Meeting of the American Educational Research Association, New Orleans, LA, United States.

Manalo, J. R., & Wolfe, E. W. (2000b, April 24–28). *The impact of composition medium on essay raters in foreign language testing* [Paper presentation]. The Annual Meeting of the American Educational Research Association, New Orleans, LA, United States.

McGuire, D. W. (1995). *A comparison of scores on the Kansas Writing Assessment for word-processed and hand-written papers of eleventh graders* [Unpublished doctoral dissertation, Kansas State University].

Mogey, N., & Hartley, J. (2013). To write or to type? The effects of handwriting and word-processing on the written style of examination essays. *Innovations in Education and Teaching International*, *50*(1), 85–93.

Mogey, N., Paterson, J., Burk, J., & Purcell, M. (2010). Typing compared with handwriting for essay examinations at university: Letting the students choose. *Research in Learning Technology*, *18*(1), 29–47.

Perelman, L. (2014). When "the state of the art" is counting words. *Assessing Writing*, *21*, 104–111.

Powers, D. E., Fowles, M. E., Farnum, M., & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement*, *31*(3), 220–233.

Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives*, *5*(3).

Russell, M., & Plati, T. (2002). Does it matter with what I write? Comparing performance on paper, computer and portable writing devices. *Current Issues in Education*, *5*.

Russell, M., & Tao, W. (2004a). Effects of handwriting and computer-print on composition scores: A follow-up to Powers, Fowles, Farnum, & Ramsey. *Practical Assessment, Research, and Evaluation*, *9*(1).

Russell, M., & Tao, W. (2004b). The influence of computer-print on rater scores. *Practical Assessment, Research, and Evaluation*, *9*(1).

Russell, M. K. (1999). *Testing on computers: A follow-up study comparing performance on computer and on paper* [Unpublished doctoral dissertation, Boston College].

Shaw, S. D. (2003). Legibility and the rating of second language writing: The effect on examiners when assessing handwritten and word-processed scripts. *Research Notes*, *11*, 7–11.

Sweedler-Brown, C. O. (1991). Computers and assessment: The effect of typing versus handwriting on the holistic scoring of essays. *Research and Teaching in Developmental Education*, *8*(1), 5–14.

Whithaus, C., Harrison, S. B., & Midyette, J. (2008). Keyboarding compared with handwriting on a high-stakes writing assessment: Student choice of composing medium, raters' perceptions, and text quality. *Assessing Writing*, *13*(1), 4–25.

Wolfe, E. W., Bolton, S., Feltovich, B., & Bangert, A. W. (1995, April 18–22). *The influence of computers on student performance on a direct writing assessment* [Paper presentation]. The Annual Meeting of the American Educational Research Association, San Francisco, CA, United States.

Wolfe, E. W., Bolton, S., Feltovich, B., & Niday, D. M. (1996). The influence of student experience with word processors on the quality of essays written for a direct writing assessment. *Assessing Writing*, *3*(2), 123–147.

Wolfe, E. W., Bolton, S., Feltovich, B., & Welch, C. (1993). *A comparison of word-processed and handwritten essays from a standardized writing assessment* (ACT Research Report Series, Issue 93-8).

Wolfe, E. W., & Manalo, J. R. (2004). Composition medium comparability in a direct writing assessment of non-native English speakers. *Language Learning & Technology*, *8*(1), 53–65.

Wolfe, E. W., & Manalo, J. R. (2005). *An investigation of the impact of composition medium on the quality of TOEFL Writing scores* (Research Report, Issue RR-72). Educational Testing Service.

Yu, L., Livingston, S. A., Larkin, K. C., & Bonett, J. (2004). *Investigating differences in examinee performance between computer-based and handwritten essays* (Research Report, Issue RR-04-18). Educational Testing Service.

Yu, W., & Iwashita, N. (2021). Comparison of test performance on paper-based testing (PBT) and computer-based testing (CBT) by English-majored undergraduate students in China. *Language Testing in Asia*, *11*(1), 32.

Zhu, Y., Shum, S.-K. M., Tse, S.-K. B., & Liu, J. J. (2016). Word-processor or pencil-and-paper? A comparison of students' writing in Chinese as a foreign language. *Computer Assisted Language Learning*, *29*(3), 596–617.

Comparing music recordings using Pairwise Comparative Judgement: Exploring the judge experience

Lucy Chambers, Emma Walland and Jo Ireland (Research Division)

Introduction

Comparative Judgement (CJ) involves judges comparing two or more artefacts (often exam responses or scripts) to decide which is better. Multiple judgements of each artefact are statistically modelled to assign each a relative measure of quality and consequently create a rank order of artefacts. CJ has been widely investigated in educational assessment as an alternative for marking (Pollitt, 2012; Steedle & Ferrara, 2016; Walland, 2022; Wheadon et al., 2020), for standard maintaining (Benton et al., 2022; Curcin et al., 2019), for monitoring comparability (Bramley, 2007; Jones et al., 2016) and, more recently, for moderation (Chambers et al., 2024; Vidal Rodeiro & Chambers, 2022).

In the context of alternatives to marking, Bramley (2022) noted in his editorial for issue 33 of *Research Matters* (which focused on CJ) that the key questions are the reliability and validity of the resulting scores, the feasibility and cost, and transparency from the candidate perspective. This article seeks to add support to the validity argument by addressing the judge perspective. It is important to verify that the judges are able to make appropriate CJ decisions just as "it is necessary to ensure that the judges themselves believe in the validity of what they are doing if stakeholders more widely are to be convinced" (Bramley, 2022, p. 7).

Decisions within a CJ context are considered to be holistic; the judges consider the evidence presented as a whole and make an evaluation. This is as opposed to the more traditional analytic method of marking using a detailed mark scheme. On the surface the CJ task appears simple, but it is actually the result of considering many pieces of interconnecting evidence. Leech and Vitello (2023) proposed three central concepts that "should define holistic judgement in an assessment context" (p. 4). Namely, the ultimate output is singular in nature, the process involves the combination of comprehensive construct-relevant evidence and that the process considers the interconnectedness of the evidence. By evaluating the judge experience, we can establish to what extent these concepts have been fulfilled.

To date, the vast majority of studies have involved written or text-based artefacts. There are a small number of studies using Art or Art and Design portfolios (Mason & Garelli, 2022; Newhouse, 2014; Tarricone & Newhouse, 2016) and one project using voice recordings (RM, 2022). To our knowledge there have been no studies involving wider-ranging artefacts, for example, recordings of music. This study sought to address this gap.

As part of a project exploring alternative ways of marking Non-Examined Assessments (NEA), we investigated using Pairwise Comparative Judgement (PCJ) to assess OCR's GCSE Music portfolios.[1] Previous work has shown that using CJ on larger bodies of NEA work (i.e., larger in size than an exam script) is practically feasible (Vidal Rodeiro & Chambers, 2022) and this study built on this by using portfolios that were primarily auditory in nature.

Previous work has also shown, however, that making comparative judgements can be challenging in certain circumstances. In a synthesis of participant questionnaires from multiple studies exploring CJ in a standard setting context, analysis has highlighted the challenges in making comparisons when the work is very different in nature (Leech & Chambers, 2022). With GCSE Music, certain differences are inherent as candidates will use different instruments, different mediums (e.g., live instrument versus sequencer) and different musical genres. In addition, pieces will be of different technical difficulty. Thus, we were keen to explore what, if any, level of challenge this might raise for the judges.

This article examines the judges' perceptions of using CJ in this context with reference to the *Dimensions of judge decision-making* model (Leech & Chambers, 2022) and makes comparisons with the findings from text-based studies.

Method

In England, OCR's GCSE Music (J536) involves one written paper (examined) and two performance-based components (Non-Examined Assessments). For the current study, we used one of the performance-based components: the integrated portfolio. This consists of a solo performance and a composition to a brief set by the candidate. The portfolios consisted of audio files, musical scores and any other accompanying documentation.

A sample of 150 NEA candidate submissions were selected from the 2019 exam series. The sample was selected using stratified random sampling based on candidate final grade. The original marks awarded by the teachers were removed, as well as any teacher commentary about how they evaluated the work. The candidate work was separated into performance and composition (so that the two elements could be judged separately) and loaded onto a bespoke online marking software. The software was user-friendly and allowed participants to listen to the audio recording (while simultaneously viewing the musical score and any other documents) and record their judgements all in one place.

1 Currently such portfolios are marked by teachers and then moderated by Awarding Organisation trained assessment specialists. For details of the process see Gill (2015).

Fifteen participants were recruited to take part in the study. They were drawn from the pool of OCR assessment specialists for GCSE Music and, as such, they were familiar with the material and assessment objectives. They were a mixture of current and retired teachers.

Each participant judged 80 pairs of performances and 80 pairs of compositions, in the order of their choosing. The pairs were determined and allocated using a randomly generated design such that each candidate's work was included in 16 comparisons. The same design was used for both the performances and compositions. Participants were instructed to choose which of each pair better demonstrated the construct of interest:

- For performances: Which student performed with better technical control, expression and interpretation (accounting for difficulty)?
- For compositions: Which student demonstrated the highest level of successful compositional skills?

Previous research (Leech & Chambers, 2022; Vidal Rodeiro & Chambers, 2022; Walland, 2022) reported that participants sometimes found it challenging to make holistic judgements and sometimes resorted to analytical marking. Thus, in this study, we enhanced the training and made specific efforts to address potential discomfort with the method. This involved familiarisation, practice, and small group online training meetings where we discussed the judgements and provided strategies to assist with decision-making. Some participants raised queries about how the method would work in practice; we asked participants to try to concentrate on the exercise and not think about the logistics. In order to mimic the support of a traditional Team Leader,[2] we supported the participants throughout the judging and offered individual online meetings to discuss any further queries.

The participants completed their judgements at their own pace, working towards a final deadline. We designed and distributed an online post-judging questionnaire where we collected participants' views and experiences of the method. Topics included likes and dislikes with the method, ease of shifting from marking to CJ, any challenging comparisons, confidence in decision-making and whether the participants found themselves re-marking or using the mark scheme.

Frequencies of responses to selected closed questions are reported alongside the question. The open-ended comments were analysed and grouped into themes that spanned across the questionnaire (i.e., the themes did not directly correspond to specific questions) – firstly, according to the *Dimensions of judge decision-making* model (Leech & Chambers, 2023) and then into other data derived themes.

When reporting results, representative comments (rather than all) are presented to capture the full breadth of opinions. Obvious typographical errors were corrected to aid readability. Px denotes the participant number.

2 A Team Leader will guide and co-ordinate a team of assistant examiners to ensure they are all marking to the same standard.

Findings

We started by asking the participants how easy they found the shift from traditional analytical marking to PCJ. Overall the participants found the shift to be straightforward saying it was "a simpler task!" (P2), that "judging quality rather than analysing criteria felt quite natural" (P7) and that it "was not looking to fit a piece into a box – just establish if it was better or worse than a second piece" (P5). Three participants were neutral and only one reported finding the shift difficult, saying that:

> "It was challenging to change to the comparisons but once I had done a few learners' work I felt more at home with it. It was a different way of addressing assessment and I did enjoy it by the end of the work" (P6)

Considering it was the first time that participants had encountered the PCJ approach, their reaction was promising.

We now look in more detail at decision-making and any challenges experienced by the participants. In order to frame the participants' perceptions of the method, we drew on the *Dimensions of judge decision-making* model (Leech & Chambers, 2023). This model (Figure 1) highlights that a judge's CJ decision-making is related to: their individual approach, the structure and features of the question paper, the way that the candidates have answered items and the unique comparative requirements of the CJ task. The arrows in the model illustrate that these dimensions impact and interplay with one another. Using this model allows us to interrogate whether judges are making appropriate decisions and therefore creating valid outputs. Table 1 summarises the judges' decision-making features found in this study. That a number of construct-relevant features are present in each dimension supports the second concept of holistic judgement specified earlier (Leech & Vitello, 2023). The sections that follow report the findings from the current study for each of the points in Table 1 in turn and, where relevant, provide reflections on how these findings compare to those from past CJ studies that involved text-based artefacts.

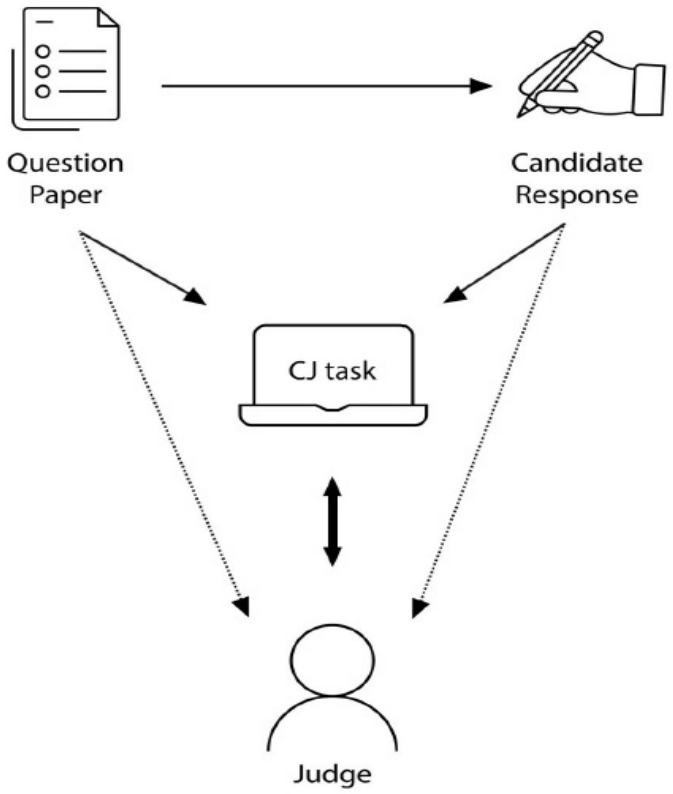

Figure 1: Dimensions of judge decision-making

Table 1: Summary of decision-making features identified in the current study by dimension

Judge-centred dimension	Question paper features dimension	Candidate response and CJ task dimensions
• Ability to make holistic judgements • Confidence • Understanding the process • Cognitive load • Judge bias	• Performance versus composition • Many pieces of information (e.g., score and recording)	• Instrument, genre/ style, medium (sequencing versus live) • Piece difficulty • Balance of different response elements • Closeness in quality

Judge-centred dimension

The first dimension of the model we will examine is the judge-centred aspect. One of the key features within this dimension is whether judges were actually able to make holistic PCJ decisions – a central tenet in ensuring the validity of the method. Whether or not participants showed any marking behaviours while conducting PCJ may be an indication of this. We found that the majority of participants reported that they never or rarely engaged in these behaviours (see Table 2). This is in line with the instructions they were given during training, which emphasised that the participants should try to avoid marking the work. Nonetheless, some participants did note that:

"I would have found it easier on several occasions to award individual marks for technical skills, expression and difficulty, and then come up with a final total to make a judgement" (P12)

"Although I found this quite easy, I did struggle with not giving pieces a mark. I had to keep mentally referring back to the old mark scheme as there is no real guidance for marking in this way" (P15)

Table 2: Participants' self-reported engagement in marking behaviour

	Never	Rarely	Sometimes	Frequently
During the PCJ part of the study, how often did you find yourself re-marking students' work (i.e., awarding marks in the traditional way)?	7	7	1	0
During the PCJ part of the study, how often did you need to refer back to the traditional mark scheme in order to make decisions?	8	3	1	3

When making holistic decisions we expect judges to draw on their experience and their knowledge of what "good" looks like and acknowledge that this may vary across judges. In the assessment context this will inevitably include knowledge of the assessment objectives and expected standards, thus reference to this would be expected. However, if these judgements become mechanistic (e.g., marking) and breach the third interconnected aspect of holistic judgement (Leech & Vitello, 2023) then the judgement is no longer holistic, which is a threat to validity. It is encouraging that marking behaviours were infrequent. In addition, the presence of some marking behaviour is not unprecedented as previous CJ studies have found that judges sometimes re-marked the work explicitly using the mark scheme or their knowledge of it (Leech & Chambers, 2022; Vidal Rodeiro & Chambers, 2022; Walland, 2022).

The self-report of the participants suggested, that for the most part, the decisions were valid. In fact, the participants noted that the exercise made them reflect on the essential features of effective performance and composition. Participant 14 noted that "it made one think harder about the fundamental principles of composing and performing to assess why one piece was better/worse than the other".

Another related feature is the judges' level of confidence in making their PCJ decisions. When asked directly, most participants reported that they were confident or very confident (see Table 3).

Table 3: Participants' self-reported confidence in their PCJ decision-making

	Very confident	Confident	Neither	Not confident	Not at all confident
How confident were you about your PCJ decisions?	2	10	3	0	0

Their comments also emphasised their confidence, for example, Participant 13 noted that "I'd say 90 per cent of the time very confident. There were just a few where I doubted my judgement" and Participant 11 reported that "generally I felt confident in the choice I made because I felt it was clear in the majority of cases". Participants cited their experience, previous marking and moderating, and their ability to play many instruments as contributing reasons for their confidence. Participant 7 gave a succinct reason for their confidence: "because it was a straight comparison of quality and musicianship". Other reasons stemmed from there often being a clear difference in quality between the pieces, and the knowledge that they were not solely responsible for the candidate's final mark. Participant 11 summed this up:

> "Some pieces were very easy to compare as the standard was so vastly different. Some were harder but I took comfort in the fact that I wasn't the only person marking the candidate so it didn't all fall on my shoulders" (P11)

One participant, who rated their confidence as "neither", reported that they found it "very difficult to compare. We are not used to doing this. We mark/moderate individuals but don't compare" (P13). This suggests that unfamiliarity may have played a part in their level of confidence.

It is also possible that the research context affected confidence levels. In fact, two participants alluded to this as increasing their confidence:

> "Actually, I felt very little pressure in doing this marking, I guess because it is a research project using 'old' candidate work. When marking/moderating 'live' work, one is much more conscious that what you do has a direct effect upon an individual's/centre's results" (P2)

> "The process has been enjoyable but I felt under no pressure of time" (P5)

This aligns with findings from the text-based standard maintaining studies cited in Leech and Chambers (2022): judges involved in live (exam session) trials of the methods found judging more challenging than those in pilot studies. Nonetheless, the high levels of confidence in PCJ found in the current study reflect those from text-based research (Vidal Rodeiro & Chambers, 2022).

Another judge-centred feature was judges' understanding of the process. Several participants wanted more information on the method – evidence on how it would work in practice (e.g., who would make judgements) and what the outcomes would be (e.g., how would final marks be derived, what feedback could be given to

schools). These points go beyond the current article's focus on judges' experiences of making judgements but have the potential to affect judges' wider confidence in the method.

In terms of the decision process, several concerns were expressed, for example, feeling bad for the losing candidate, having to make a judgement when the participants felt the pieces were of the same standard, or not seeing the benefits of the method:

> "A lot were very easy, but sometimes I liked both or thought both were not so good. Sometimes there was a very good performance and an exceptional performance and I felt bad saying the exceptional one was better, when the very good one would have been the best in many other pairings" (P8)

> "Very often there was a distinct difference between the two pieces being listened to. I was just a little uncomfortable marking one piece as being better than another when they were of the same standard (especially at the top end)" (P15)

> "I'm not sure what the gains would be or what would be achieved beyond the traditional methods unless comparisons were made between pieces of a similar type. Even then how would you compare a rock singer with a more classically trained singer" (P14)

> "I found it straightforward to shift but I'm not confident about the results it will produce, even when all the moderators' decisions are put together, some decisions could have gone either way. I think the top and bottom candidates will be in the right place but I'm not sure about all the ones somewhere in the middle" (P9)

Related to understanding the process, some judges commented on the method itself:

> "I can see the benefits of the PCJ method and I believe that if moderators are trained to complete this approach it would be successful. I would think that moderators would listen to more pieces of music which again would be a good thing" (P6)

> "This method is very subjective" (P8)

> "It just felt a bit random to me. It didn't seem like I was rewarding the candidate's work" (P9)

In previous research on text-based studies there has often been one or two judges who did not favour CJ as a method (Vidal Rodeiro & Chambers, 2022; Walland, 2022), so it is not surprising that some caution about how the method worked was expressed by some of the current participants.

In terms of cognitive load, the pieces of music were often quite long, and some participants struggled with remembering the first piece after listening to the second, for example, Participant 13 said that it was "too long after listening to both examples to remember the first one sufficiently". Participants reported that they sometimes took notes as a memory aid. Participant 1 noted that "listening to music takes time! Notes needed to be taken in order to remember back to piece 1". Also related to note taking, Participant 3 commented that "it became rather dull in places as a lack of marking / note writing to help lead to a conclusion led to a lack of brain power / interest at times".

The cognitive load needed to complete the activity has been discussed in other text-based studies (Vidal Rodeiro & Chambers, 2022; Walland, 2022). Interestingly, the challenge noted here, of recalling the first artefact, was not apparent in the text-based tasks, as a quick view or skim of the first text-based script would be enough for the judge to recall the content – with music, there is an absence of such cues.

Some participants mentioned judge bias as a feature:

> "I found judging drummers very hard with other performances and I wonder if I was harsher there on the drummers" (P11)

> "No real dislikes – sometimes a close call was hard to make. Possible scope for bias by the assessor against work in certain genres, meaning that the wrong piece is preferred...?" (P7)

> "In a real situation I feel judgement could be clouded at times when hearing something new or refreshing i.e., a steel pan after listening to 3 or 4 vocal pieces in a row" (P3)

This is an interesting finding since judge bias has not been previously raised by participants in text-based studies.

Overall, the judges felt able to and were confident in making judgements. However, similarly to text-based studies (Leech & Chambers, 2022; Vidal Rodeiro & Chambers, 2022), the participants did experience challenges in making the judgements due to the interplay with other dimensions. The next sections discuss the other dimensions.

Question paper features dimension

For GCSE Music NEA, there is no question paper as such. However, candidates produce a recorded performance and performed composition, so we can think of these as essentially two items, weighted equally. The participants found that compositions appeared to present more problems than performances. Participant 6 noted that "the performances were more straight forward". Other participants also reported this and added additional detail about the interaction with medium and cognitive challenge:

> "I found performances easier than compositions to compare especially if it was a live composition played well compared to a computer export" (P11)

> "Composition required a consideration of the whole piece more so than performances" (P7)

There were often many pieces of information, for example, cover sheets, musical scores and the candidate recording. The interface of the software was designed to be user-friendly, however navigating through this work and viewing it clearly was sometimes a challenge for judges. One participant noted that "some candidates had about 60 pages of score" (P1). Another noted that:

> "I would have liked to be able to jump between documents. There were numerous occasions where I would have liked to have jumped to a cover sheet, which was the final document, but I had to scroll through page after page of score to get to it. Also, a zoom function would have helped at times" (P12)

A related issue has previously been found with text portfolios, where participants experienced some difficulties when scrolling through many pages of work due to time lags (Vidal Rodeiro & Chambers, 2022) and difficulty making decisions due to the layout of portfolios.

We found that the features apparent for this dimension were quite different to text-based judgements, due in part to the absence of a question paper containing discrete items. For text-based tasks, the features mentioned by judges were: number of short items, the presence of longer questions involving evaluation or explanation and the focus on more discriminating items over others (Leech & Chambers, 2022).

Candidate response and CJ task features dimension

The candidate response features mentioned by participants included elements such as instrument, genre and style, difficulty of piece and medium. We found discussion of these features to be inextricably bound with discussion of the PCJ task. Comments centred around balancing the different response features when making comparisons between the candidates.[3] As a result, we discuss both dimensions together.

Participants reported that for the most part the decisions were straightforward, and that "most of the time there were few problems differentiating pieces" (P14). However, when the pieces were very different in some way – for example, "perhaps one was technically accurate but emotionless, another full of expression but out of tune" (P9), or "a difficult piece played badly with an easier piece played really well" (P10) – then comparison could be more challenging. Interestingly, participants

3 This may be in part due to the nature of the survey question. In this study we asked a question about whether they found any comparisons challenging rather than an explicit question on how the participants made their decisions.

differed in what they found challenging. Table 4 highlights some of the response elements and the differing views.

Table 4: Differing participant views (quotations) with respect to candidate response features

Response feature	Perspective – easy	Perspective – neutral	Perspective – challenging
Difficulty of musical piece	It was easy to compare performances where the difficulty level was different. It was much harder to compare performances which were very similar in standard (P8)	Overall, regardless of the instrument, there were several performances that were difficult to determine which was better and sometimes it was the difficulty of the piece that was the decider (P3)	The biggest challenge for me was comparing pieces with widely different difficulties. There were easy pieces that were played fluently and with style, compared with significantly harder performances that had hesitations, etc. (P12)
Instruments	… I found it okay to compare performances on different instruments (P8)		Difficult when marking completely different instruments i.e., Piano versus Indian Raga vocal line (P3)
Genres/styles	I actually found it quite straightforward to compare a range of different genres. The quality of a great composition or performance shone through regardless of the genre (P7)	I think it is always hard to mark things that one is less familiar with such as classical Indian music or sequencing (P4)	It was sometimes difficult when marking the same instrument which were similar in credit but of different styles i.e., a Big Band drummer playing live versus a Grade 8 Rock drummer (P3)
Medium (sequencing versus live)			Sequencing against "live" instrument was difficult. … (P5)

Interestingly, these features seemed to have more impact on participant comments than some of the features to be assessed as set out in the mark scheme (e.g., for performance: technical control and fluency and expression and interpretation; and for composition: sense of style, a range of musical elements, composition techniques, stylistic and structural conventions). This could be evidence of these response features getting in the way or perhaps evidence of the participants judging holistically.

Several participants reported that it was challenging to judge between two candidates whose work was very similar in quality: Participant 11 reported that "very occasionally I wanted to say it was a tie as I really felt both pieces were the same standard", Participant 15 cited instances "where the same mark would have been awarded to both in the usual mark scheme" and went on to report that "when work was of an identical standard there was no option to show this - you still had to choose which one was better". This is a challenge that has been seen across previous CJ studies. Judges often struggle in this scenario as it goes against their many years of training and their wish to do right by the candidate. In the training as part of the current study, we tried to reassure the participants and explained that the method, with multiple judgements, would ensure the appropriate outcome for the candidate. The fact that participants worried about this issue despite the training suggests that further reassurance and evidence needs to be provided to judges (and other stakeholders).

The participants reported a number of strategies for dealing with the challenge of comparing work of similar quality:

> "In most cases one candidate's work seemed clearly better than the other. When this was not the case I made my best judgement and trusted that the system would work" (P4)

> "Where there were close calls, it was back to basics – who was the most accurate and the most musical and which piece was delivered the most successfully given the challenge of the repertoire" (P7)

> "With some less able musicians it was sometimes a case of which one was worse rather than better and working it out that way" (P5)

> "Another challenging performance was a Rap artist whose performance was stylish and professional versus an alto sax performance. I found myself taking other things into account opting for the sax as this candidate would have had to learn how to play the instrument and follow the music over a longer period of time" (P3)

This last comment shows how other, potentially unintended, factors might be used where judgements are difficult. Some participants' comments showed their awareness of the need to know the criteria to be used even when making comparative judgements:

> "It is easier to compare 2 pieces rather than trying to fit them into a level category. You still need to know/understand the criteria on which you are judging the pieces" (P5)

The features described in this dimension were again often different from those found in text-based studies. For text-based studies, candidate response features were centred around response consistency, depth of responses, clarity/structure,

spiky profiles[4] and omitted questions, and use of examples, facts and statistics (Leech & Chambers, 2022; Vidal Rodeiro & Chambers, 2022; Walland, 2022). For Music, as there was only one non-text task in each condition, some of these responses were not present (e.g., use of examples, facts and statistics) or were presented differently (e.g., an imbalanced performance instead of a spiky profile).

For GCSE Music there are many variables (e.g., instrument, medium, difficulty and genres, etc.) and it appears that it is the interactions between these features and the many permutations and combinations that prove challenging.

Fairness

Moving beyond the dimensions model, another related theme that came up in the responses was that of fairness. One participant stated that "it just doesn't seem very fair, the two being compared are so different, e.g., a big band composition on Sibelius compared to a garage band piece, or a film music composition compared to a piano piece" (P9) and "it would be fairer to compare similar instruments where possible" (P9).

Another concern was that candidates would not receive a fair grade, as Participant 3 noted:

> "I didn't enjoy this method. It felt less personal and less hands on with a lack of professional opinion. I felt that in some cases, there wasn't a need for expertise or musicianship to be able to determine 'which was better' and that the candidates would not receive a fair and considered grade." (P3)

In contrast, some participants saw the inherent fairness in the method itself due to multiple judgements, for example, "it felt fairer that the marks would be based on lots of people's opinions" (P4) and "I guess the more times a candidate's work is viewed by different assessors, the more chance there'll be of establishing a true and fair assessment" (P2). Fairness was also cited in comparison to the current moderation process:

> "It appears to be a fairer system of marking. Although it is still subjective, the fact that a number of people would mark the same pieces should make for a better consensus. It would no longer be the school's opinion versus the (single) moderator's opinion" (P5)

> "A range of markers look at work from a range of centres, so one marker is not responsible for marking all the work of one centre – this provides a balance of opinion" (P7)

> "I think centres would welcome the idea that the work is marked multiple times to establish a clear overview of the relative standard of the work" (P7)

The current finding regarding the benefit to fairness of multiple judges evaluating

4 A spiky profile is where candidates answer some questions well and others poorly.

one candidate's work echoes views reported in a text-based GCSE English Language CJ study (Walland, 2022).

Conclusion

This study sought to investigate, from the judge perspective, the use of PCJ with auditory-based artefacts. In this study we used GCSE Music, and as such it is important to note that the findings as detailed relate to music recordings and will not necessarily apply to all auditory-based artefacts. This study used a small sample of participants (n=15) and one component of GCSE Music. It involved only self-report data; observational research could add richness and support to the findings. As such, these factors should be borne in mind if generalising the findings more broadly.

The use of auditory-based artefacts, in particular music files (as in this study), is an under researched context for CJ. At the start of this article, we noted two aspects of the judge experience that are necessary to support the validity of the CJ method. Namely, whether judges are able to make appropriate CJ decisions and whether they believe in the validity of what they are doing.

The enhanced training, familiarisation activities and support we gave participants appeared to have proved effective. We saw that for the most part judges were able to make appropriate decisions, there was little evidence of participants re-marking or attending to construct-irrelevant features, and the judgements involved the balancing of different response elements. This also suggests that the second and third aspects of holistic judgement, as defined by Leech and Vitello (2023) (that comprehensive relevant evidence is used and interconnectedness is considered), were met. (Note that Leech and Vitello's first criterion of holistic judgement is also met, since the participants provided a singular judgement for each pair of performances or compositions.)

In terms of whether the judges believed in the validity of what they were doing, the findings were mixed. Participants could see the benefits of having multiple judgements of each candidate's work and there was also some evidence that participants were revisiting the fundamental principles of composing and performing. Some participants, however, appeared unconvinced by the method. Sometimes it was a lack of understanding or belief in the process – this was particularly for work they considered to be of the same standard. Further training, experience and provision of evidence could help alleviate this.

A key concern related to candidate work that was very different in some way – for example utilising different genres or instruments, or when the piece difficulty varied. This is harder to address. Leech and Chambers (2022) noted that in the CJ context "there is no immediately clear way to determine which paper of a pair or pack is the superior if each is better in a different way" (p. 45). They discussed the tension between the way current exam papers are set up (i.e., to be marked) and holistic CJ judgement which relies on a judge's conception of what constitutes better performance. Leech and Vitello (2023) described this as an "informal rubric" where judges determine which features to prioritise. They

"contend this informal rubric should be made more formal by the provision of more explicit guidance, and the comparison simplified by, if at all possible, ensuring the similarity of form between different artefacts" (p. 18).

In this study we did provide some guidance, however for the participants it was the first time they had used this method, and it was unsurprising that some challenges remained. It is recommended that similar guidance and training should accompany further CJ studies so that judges feel confident in making independent holistic decisions and are clear which elements would be considered construct irrelevant in any context. This would help ensure the validity of assessment outcomes.

In terms of simplifying the comparison, for GCSE Music, pairing similar artefacts would be practically unfeasible. Even if, for example, pieces were paired on one factor such as instrument, the genre can be vastly different. However, further research utilising observation-based methods could be used to render the methods by which judges resolve this challenge explicit. In parallel, specific research into the effects of instrument and genre on CJ outcomes could also be conducted to explore whether any bias exists.

What these challenges show is the complexity of making CJ decisions – far from an instant decision, a holistic judgement is the "consequence of the aggregation of a series of micro-judgements, each of which might be quite different for each judge making them" (Leech & Vitello, 2023, p. 13). The level of challenge can be further increased when an element of "difference" is added. All artefacts involving some level of candidate choice, whether text or auditory-based, will create challenges for CJ as difference will be inherent. This difference could be for example, topic in History or choice of sport in PE. Music raises this level of challenge further in that so many elements interplay with each other. It is possible that there could be a "difference ceiling" – a point in certain contexts where the artefacts are just too different to be compared validly using CJ, and other methods such as analytic marking or "levels-only" marking would be more suitable (for information on "levels-only" marking see Walland and Benton, 2023).

Some of the challenges the participants experienced were more practical in nature, for example, the cognitive load in remembering the first artefact or ease of viewing any documents while listening to the recording. These factors are unlikely to be restricted to music recordings and could apply to other portfolios containing audio recordings. Care should be given with respect to the length of any recordings. If portfolios containing large quantities of evidence are to be used alongside audio recordings, then it is necessary to consider which pieces of evidence should be included. Clear design and user experience testing of any software are vital.

It is important to note that, for the most part, participants found the shift to PCJ straightforward and felt confident making the judgements. What was particularly clear from this study was that, in general, participants were open to new ideas and ways of working and welcomed the opportunity to be involved in the research.

References

Benton, T., Gill, T., Hughes, S., & Leech, T. (2022). A summary of OCR's pilots of the use of comparative judgement in setting grade boundaries. *Research Matters: A Cambridge University Press & Assessment publication, 33,* 10–30.

Bramley, T. (2007). Paired comparison methods. In P. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 246–300). QCA.

Bramley, T. (2022). Editorial – the CJ landscape. *Research Matters: A Cambridge University Press & Assessment publication, 33*, 5–8.

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, 31*(1), 32–55

Curcin, M., Howard, E., Sully, K., & Black, B. (2019). Improving awarding: 2018/2019 pilots. *Research Report Ofqual 19/6575. Research and Analysis*.

Gill, T. (2015). The moderation of coursework and controlled assessment: A summary. *Research Matters: A Cambridge Assessment Publication, 19*, 26–31.

Jones, I., Wheadon, C., Humphries, S., & Inglis, M. (2016). Fifty years of A-level mathematics: have standards changed? *British Educational Research Journal, 42*(4), 543–560.

Leech, T., & Chambers, L. (2022). How do judges in Comparative Judgement exercises make their judgements? *Research Matters: A Cambridge University Press & Assessment publication, 33*, 31–47.

Leech, T., & Vitello, S. (2023). What is a holistic judgement, anyway? *Research Papers in Education*, 1–23.

Mason, K., & Garelli, L. (2022). *Assessment of art and design courses using comparative judgement in Mexico and England.* Paper presented at the annual conference of AEA-Europe, Dublin, November 2022.

Newhouse, C. P. (2014). Using digital representations of practical production work for summative assessment. *Assessment in Education: Principles, Policy & Practice, 21*(2), 205–220.

Pollitt, A. (2012). Comparative judgement for assessment. *International Journal of Technology and Design Education, 22*(2), 157–170.

RM. (2022). *Using Adaptive Comparative Judgement as a reliable way to assess oracy at scale*.

Steedle, J. T., & Ferrara, S. (2016). Evaluating comparative judgment as an approach to essay scoring. *Applied Measurement in Education, 29*(3), 211–223.

Tarricone, P., & Newhouse, C. P. (2016). A study of the use of pairwise comparison in the context of social online moderation. *The Australian Educational Researcher, 43*, 273–288.

Vidal Rodeiro, C., & Chambers, L. (2022). Moderation of non-exam assessments: Is Comparative Judgement a practical alternative? *Research Matters: A Cambridge University Press & Assessment publication, 33*, 100–119.

Walland, E. (2022). Judges' views on pairwise Comparative Judgement and Rank Ordering as alternatives to analytical essay marking. *Research Matters: A Cambridge University Press & Assessment publication, 33*, 48–67.

Walland, E., & Benton, T. (2023). Multiple marking methods as alternatives to single marker analytical essay marking: Exploring pairwise comparative judgement, rank ordering and levels-only [Manuscript submitted for publication].

Wheadon, C., Barmby, P., Christodoulou, D., & Henderson, B. (2020). A comparative judgement approach to the large-scale assessment of primary writing in England. *Assessment in Education: Principles, Policy & Practice, 27*(1), 46–64.

Research News

Lisa Bowett (Research Division)

The following reports and articles have been published since *Research Matters*, Issue 37:

Journal articles and other publications

Constantinou, F. (2024). Assessing students' application skills through contextualized tasks: Toward a more comprehensive framework for embedding test questions in context. *Practical Assessment, Research, & Evaluation, 29*(10). https://doi.org/10.7275/pare.2103

Crisp, V., Elliott, G., Walland, E., & Chambers, L. (2024) A structured discussion of the fairness of GCSE and A level grades in England in summer 2020 and 2021. *Research Papers in Education*. https://doi.org/10.1080/02671522.2024.2318046

Gill, T. (2024). Core Maths qualifications: how they fit in post-16 programmes of study and their impact on other subjects with a quantitative element.

Gill, T. (2024). Is Core Maths fulfilling its aim? Impact on higher education outcomes.

Johnson, M., & Majewska, D. (2024). What is non-formal learning (and how do we know it when we see it)? A pilot study report. *Discover Education* 3(148). https://doi.org/10.1007/s44217-024-00255-y

Kreijkes, P. (2022). A Bird's-Eye View of Curriculum Publications Concerning Seven Countries: A Bibliometric Analysis. ISSN: 2188-1162. *The European Conference on Education 2022: Official Conference Proceedings*. https://doi.org/10.22492/issn.2188-1162.2022.29

Majewska, D., & Johnson, M. (2024). Uncovering the landscape of cross-national UK education research: An exploratory review. *Educational Research, 66*(2), 205-227 https://doi.org/10.1080/00131881.2024.2334751

Oates, T. (2024). The COVID-19 pandemic may be a thing of the past – its impact in schools is not. Covered by TES, The Telegraph, The Guardian, The Times, Schools Week, School Management Plus, Education Journal, together with a wide variety of online and print sources across Europe, and North and South America.

Vidal Rodeiro, C. L. (2024). Progression of the 2020 Key Stage 4 cohort to post-16 study.

Conference presentations

Abu Sitta, F., Maddox, B., Casebourne, I., Hughes, S., Kuvalja, M., Hannam, J., & Oates, T. (2023). The Futures of Assessments: Navigating Uncertainties through the Lenses of Anticipatory Thinking. Cambridge Assessment Network Conference, Cambridge, UK. (17 April) – Sarah Hughes presented.

Carroll, M. & Constantinou, F. (2024). Teachers' experiences of teaching during the Covid-19 pandemic and some positive implications. ResearchED Conference, Cambridge, UK. (20 April)

Elliott, G., Rushton, N., & Ireland, J. (2024). Is the GCSE incongruous in the light of other jurisdictions' approaches to assessment? Cambridge Assessment Network Conference, Cambridge, UK. (17 April) – poster reused for this event.

Gill, T. (2024). Research into the potential benefits of taking the EPQ for concurrent and future attainment. EPQ teachers' conference (Southampton University, 11 June)

Greatorex, J., & Ireland, J. (2024). Comparing curricula from different regions: a common practice revamped by using MAXQDA. Cambridge Assessment Network Conference, Cambridge, UK. (17 April) – poster reused for this event.

Oates, T. (2024). Research around the initial acquisition and later development of reading. ResearchED Conference, Cambridge, UK. (20 April)

Oates, T. (2024). The many meanings of personalisation of learning - the good the bad and the ugly. ResearchED Conference, Cambridge, UK. (20 April)

Oates, T. (2024). The importance of strategic planning and aligned actions to raise the quality of education. Presentation in Porto, Portugal. (14 May)

Oates, T., & Suto, I. (2024). Knowledge and skills – knowledge versus skills – false oppositions and fallout. AEA-Europe Holistic Assessment SIG webinar. (13 June)

Rushton, N. (2024). Timeline of changes to the national curriculum, Cambridge Assessment Network Conference, Cambridge, UK. (17 April) – poster reused for this event.

The 2024 British Educational Research Association (BERA) conference took place in Manchester on 8 - 12 September, https://www.bera.ac.uk/conference/bera-conference-2024-and-wera-focal-meeting. Our researchers presented three papers:

Greatorex, G. Indigenous Knowledges in school curricula: a literature review and document analysis (co-researched with Jo Ireland)

Constantinou, F. Synchronous Hybrid Teaching: A More Flexible and Inclusive Mode of School Instruction?

Lestari, S. Typing Versus Handwriting Exam Scripts: Evidence Synthesis and Implications for Practice and Research

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, 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)

CAMBRIDGE

Transform your understanding of assessment

with a Cambridge postgraduate qualification

The **Postgraduate Advanced Certificate in Educational Studies: Educational Assessment** is a 15 month, part time online qualification run in partnership by the University of Cambridge Faculty of Education and Cambridge Assessment Network.

Worth 90 credits at Master's level (Level 7), this practice-based qualification will teach you to apply research methodologies to your professional context.

Starting September 2025

Find out more
cambridgeassessment.org.uk/pgca

CAMBRIDGE

The Assessment Network

Join our network

The Assessment Network is a global leader in professional development, covering assessment principles, practices and insights.

Enhance your status as a recognised assessment expert by joining a global network of assessment professionals.

For more details visit
cambridgeassessment.org.uk/the-network

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

The latest product from Cambridge Mathematics

CoffeePods are an audio-visual format of our existing Espressos, two-page filtered research summaries on a mathematics education topic.

Each CoffeePod is approximately 20 minutes long, and has a downloadable transcript, a separate PDF of the references and a link to its original matching Espresso.

So far in the series:

- EAL (English as an Additional Language) students in the mathematics classroom
- Factors, multiples and prime numbers
- The number line

Listen now

CAMBRIDGE
Mathematics

Contents / Issue 38 / Autumn 2024

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

ResearchDivision@cambridge.org
www.cambridge.org