CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Research Matters

Issue 41 / April 2026

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: Lim, C. H. J., & Kreijkes, P. (2026). Gaming and social media browsing: Evidence of links to wellbeing among girls and boys based on data from PISA 2022. *Research Matters: A Cambridge University Press & Assessment publication, 41,* 9–31. https://doi.org/10.17863/CAM.127730

Credits

Reviewers: Aylin Albayrak Sari, Carmen Vidal Rodeiro, Elinor Lieber, Filio Constantinou and Alejandra Miranda

Editorial and production management: Victoria Crisp, Lisa Bowett and Jackie White

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 41 DOI: https://doi.org/10.17863/CAM.128122

All details are correct at the time of publication in April 2026.

Contents

Foreword

Jane Mann

There are few responsibilities more important than the nurture and development of the next generation. Education shapes almost every aspect of learners' lives – where they spend much of their time, what they learn to value, the activities they undertake, friendships they forge and the strategies they develop to navigate challenge and change. All these experiences influence not only academic achievement, but the foundations of lifelong success.

Across the world, care for wellbeing is at the heart of positive, supportive learning environments. Understanding the impact that choices made at school have on long-term success, and what activities outside school, such as gaming, provide in terms of benefit and challenges, has direct relevance for school policy-making. Equally, perceptions of fairness in assessment, from the different designs to the scheduling of examinations, plays a crucial role in building trust and confidence in the system among learners.

Every article in this edition of Research Matters speaks, in one way or another, to the wellbeing of students and their educators. Reading them has made me reflect on the centrality of wellbeing to the core purpose of education, enabling learners to develop the knowledge, skills and dispositions that will help them thrive throughout their lives. Each of us has a part to play in the positioning of wellbeing in the learning environment and in fostering the understanding that what matters to learners matters to everyone.

Editorial

Victoria Crisp

Welcome to the April 2026 issue of *Research Matters*. You will notice that we have added a new section called *In This Issue*. Feedback welcome!

A learner's wellbeing can influence their educational success. Therefore, it is important for us to understand the factors that can influence wellbeing. In our first article in this issue, Carmen Lim and Pia Kreijkes explore links between amount of time spent gaming or browsing social media, and different aspects of wellbeing. Their analysis used questionnaire data from PISA 2022 to investigate associations and found some different patterns for boys and girls.

Our second article, by Tim Gill, reports an investigation of whether exam scheduling influences student performance in A Level exams. Using exam data, Tim compiled several measures such as the number of days since a candidate's previous exam and the number of exams taken so far in the exam series. Regression analysis was used to look for any relationship with exam performance, while controlling for other factors such as gender, centre type and wider attainment.

There is growing interest in increasing the use of digital exams in general qualifications. However, such change is likely to be gradual, meaning that both digital and paper-based exam routes may be required – and that care will be needed to ensure comparability between exam modes. In our third article, Carmen Lim, Carmen Vidal Rodeiro and Tim Gill use data from Cambridge's Digital Mocks Service and from live paper-based exams to explore differences in the difficulty of the same questions across modes. They also investigated whether particular characteristics of questions might influence the likelihood of differences.

In our fourth article, Annaliese Micallef Grimaud reviews the literature on factors that predict success in vocational education and training. The literature review is organised by types of predictors: prior attainment; cognitive abilities (such as verbal or numerical abilities); and non-cognitive traits (such as facets of personality). The apparent relationships between such factors and the choice to take vocational qualifications are also discussed.

Our issue finishes with an article focused on teachers' assessment literacy, by Victoria Coleman and Martin Johnson. Their research involved a questionnaire survey to gather views from teachers with professional experience of examining. The questionnaire explored their understandings of the term "assessment literacy", the perceived influence of their training, teaching and examining on their own assessment literacy, and the specific influence of examining on their assessment knowledge, skills and beliefs, and on their teaching practice.

In This Issue

Victoria Crisp

This section aims to provide a quick introduction to the research *In This Issue* including some context, key findings and their relevance.

Social media use affects some aspects of wellbeing but not others

The recent ban on social media use for under 16s in Australia and discussions of restrictions in other countries suggest growing concern about its impact on young people. One worry is the possible effect on wellbeing – however, research findings about the impact of digital use on adolescent wellbeing are often mixed or inconsistent, making it difficult to know how to interpret the evidence.

In this issue, Carmen Lim and Pia Kreijkes unpick the apparent effect of social media browsing (and, additionally, gaming) on different aspects of wellbeing, providing more detailed insights which could help inform this debate. Their findings come from analysing questionnaire data collected in an international survey of adolescents. The analyses revealed that more time browsing social media was associated with increased reports of symptoms such as anxiety, sleep difficulties and headaches and, for girls, increased likelihood of poorer body image perception, general wellbeing and life satisfaction. Moderate use of social media had some benefit for sense of belonging at school, but this positive effect was relatively weak.

Overall, the research findings could add support to arguments for restricting social media use for children and adolescents. How policy could tackle the gender-based impact on wellbeing while not introducing inequality is more difficult. This may require efforts to address certain aspects of social media content or advertising that may be contributing to specific effects on body image, for example.

Carmen and Pia's research study highlights the importance of disaggregating different aspects of wellbeing, and exploring patterns for girls and boys, in future research in this area.

Exam timetabling less important than might be expected

Students, teachers and parents may worry about whether the timetabling of exams will influence how well students perform. Do candidates suffer from fatigue as they take more exams or do they get better at later exams as they more get into an exam-taking mindset? Does it have an impact if there is little time for revision between exams?

For a factor that could potentially affect many exam takers in many different exam contexts, there is surprisingly little research about the scheduling of exams. In this issue, Tim Gill shows that, for A Levels, the effects of timetabling on exam performance are minimal, with little systematic influence except where learners take two (or more) exams on one day. There was, however, an interesting pattern in terms of ability – higher ability students seemed to benefit from having taken more exams already during the exam series, suggesting a positive effect of practice, while lower ability students seemed to be affected by fatigue.

The findings offer reassurance that the timetable of a particular student's A Level exams is unlikely to substantially affect how well they perform, and Tim concludes that there is not a strong case for substantial changes to current exam scheduling practices.

It is interesting to consider whether there are ways that lower ability students can be supported to reduce the possible impact of fatigue as the exam series progresses. For example, could additional revision lessons during the exam series help with motivation and counter the fatigue effect? In England, there are currently recommendations (from the Curriculum and Assessment Review report[1]) to reduce the amount of exam time for GCSEs (taken at around age 16 years) by at least 10 per cent. It would be informative to explore whether the ability-based pattern observed in the current analysis of A Levels is also present for GCSE students and whether the proposed reduction in exam time changes this.

Paper versus computer: Ensuring fairness by exploring differences

The third article in this issue highlights the importance of being guided by evidence when developing and delivering new assessment formats. As the use of digital exams is set to increase in the coming years, this type of research into paper-versus-digital formats is critical.

Carmen Lim, together with Carmen Vidal Rodeiro and Tim Gill, report on a study exploring whether individual exam questions differ in difficulty when delivered on paper compared with on screen. Using data from digital mock exams and paper-based live exams, they found that about one in five questions were easier in one mode than the other and that there was no clear consensus on whether digital or paper-based assessment was easier. Research in other exam contexts has also reported differences in difficulty for some questions but, as the authors discuss, there is currently limited guidance on what level of variation should be considered problematic or warrant further investigation.

1 https://www.gov.uk/government/publications/curriculum-and-assessment-review-final-report

The researchers also explored whether particular features of questions could explain the differences in performance between modes but found very few systematic associations.

Decisions about vocational courses should be informed by multiple factors

Predicting how well a student is likely to do in a particular course is tricky but important. If a student is accepted to a course that turns out not to be well suited to them, this can result in weak performance or dropping out, both of which can set back their educational or employment progress. In England, plans to shake up Level 3 vocational education by replacing existing qualifications with new V Levels will affect students' choices, making it a pertinent time to reflect on what is known about factors that can predict success in vocational education and training.

For her article in this issue of *Research Matters*, Annaliese Micallef Grimaud reviewed UK and international research literature in this area. The evidence showed that previous exam results or school grades were often good predictors of success in vocational training and qualifications. Additionally, certain cognitive abilities and other non-cognitive traits may play a role. Understanding the factors that predict success can help students decide on an appropriate educational route for them and help admissions staff select which students to admit to courses. Such understanding could also help with designing assessments to support such decision-making.

Examining benefits teaching

Whether initial training courses provide future teachers with sufficient knowledge and skills about assessment is an area of debate. Given the many competing demands on the knowledge and skills that need to be covered in teacher training, perhaps there is limited scope for additional content on assessment. However, research by Victoria Coleman and Martin Johnson shows that working as a professional examiner may well have benefits for teacher assessment literacy that extend into the classroom.

Victoria and Martin designed and administered a questionnaire to teachers who also worked as professional examiners. Most respondents felt that their assessment literacy was significantly influenced by their examining work. The benefits related to increased understanding of summative assessment, including exams, as well as formative aspects of assessment such as giving feedback to students – thus, feeding into teachers' classroom practice more broadly than just in terms of preparing students for exams. Teachers also reported that they felt more trusted by students and by colleagues on matters of assessment because of their examining experience.

School senior leaders could consider whether teacher involvement in examining is something they wish to actively encourage.

Gaming and social media browsing: Evidence of links to wellbeing among girls and boys based on data from PISA 2022

Carmen H. J. Lim (Research Division) **and Pia Kreijkes** (Research Division)

Introduction

Adolescent wellbeing is on a declining trajectory. Adolescents today are less satisfied with their life compared to adolescents two decades ago. This decline in life satisfaction was reported in the 2024 World Happiness Report among young people aged 15 to 24 in regions such as Western Europe, South Asia, Middle East and North Africa, and North America, Australia and New Zealand (Marquez et al., 2024). Similarly, when asked about their happiness in five life aspects, adolescents aged 10 to 15 in the United Kingdom (UK) reported being least happy with their appearance, school and schoolwork, followed by friends and family – all of which were on a falling trend over the past decade (The Children's Society, 2025).

Feeling emotionally well and being able to function effectively are essential for adolescents to thrive in school and beyond. Adolescents who struggle emotionally or functionally may skip school more often, lose focus in class, or show less interest in learning (Gutman & Vorhaus, 2012). Prolonged periods of low wellbeing can also increase the risk of developing mental health conditions, such as clinical depression (World Health Organization, 2025). These occurrences negatively affect adolescents' learning and academic performance, preventing them from reaching their full potential (Kaya & Erdem, 2021).

There are a multitude of factors that may cause low wellbeing among adolescents. These include financial challenges, lack of family support and friends, or being bullied at school, as reported by UK adolescents in the 2025 *Good Childhood Report* (The Children's Society, 2025). The report also concluded that heavy social media use can have a detrimental effect on adolescents' wellbeing.

Parents and policymakers are concerned about adolescents' exposure to harmful content and cyberbullying on social media and other online platforms, as well as the impact this has on their wellbeing. These concerns have prompted ongoing efforts to tighten regulations aimed at improving online safety for children and adolescents, such as the enactment of the Online Safety Act in the UK in 2023.[1]

1 For more, see https://www.legislation.gov.uk/ukpga/2023/50.

Furthermore, discussions and proposals to raise the minimum age for social media use beyond 13 years are gaining momentum globally, with Australia having legislated a ban for users under 16 and Denmark for those under 15 (Keaten, 2025).

The focus on social media in these debates is perhaps not ungrounded, considering that adolescents spend a significant amount of time on social media platforms. A study tracking detailed smartphone usage among adolescents aged 11 to 17 in the United States (US) found that social media applications were used the most each day, followed by YouTube and gaming (Radesky et al., 2023). Another survey of US adolescents indicated that they spent an average of about 2.9 hours daily on social media and an additional 1.9 hours on YouTube, totalling 4.8 hours across both activities (Rothwell, 2023). This represents roughly one-third of a 16-hour waking day.

One concern about adolescents spending such a high number of hours on their smartphones (or other digital devices) is that it may come at the expense of time spent on other beneficial activities, such as exercising, in-person play, and reading. Another concern is that heavy users are more likely to develop problematic behaviours and addiction-like symptoms, such as an inability to control usage. The Health Behaviour in School-aged Children (HBSC) study by the World Health Organization found that more than one in ten adolescents showed signs of problematic social media use or problematic gaming behaviour (Boniel-Nissim et al., 2024). Interestingly, the same study showed that more girls exhibited signs of problematic social media use than boys (13 per cent versus 9 per cent), whereas more boys showed signs of problematic gaming behaviours (16 per cent versus 7 per cent).

Various studies have highlighted differences in usage patterns between girls and boys, as well as potentially varying effects on their wellbeing. According to the HBSC study, at least 40 per cent of boys reported daily gaming, while this ranged from 15 per cent to 25 per cent for girls, depending on age. By contrast, more girls, older girls in particular, reported being in contact with friends and others through social media almost all the time throughout the day (Boniel-Nissim et al., 2024). Research also found that girls are more at risk of the negative impacts of social media use, particularly of experiences of depressive symptoms, anxiety and internalising problems (Fumagalli et al., 2024; Svensson et al., 2022). Girls also tend to be impacted at an earlier age than boys (Orben et al., 2022).

However, when examining the literature on digital use and adolescent wellbeing more broadly, researchers conclude that the findings appear mixed and inconsistent (Valkenburg et al., 2022). Some key issues contributing to these inconsistencies are that many studies do not analyse results by gender and are not specific about the measures of digital use, e.g., some consider screen time, while others look at specific digital media activities (Svensson et al., 2022).

Additionally, wellbeing is a multifaceted concept and does not have a single agreed-upon definition (McLellan & Steward, 2015). Different studies have used different indicators to measure adolescent wellbeing, such as general satisfaction with their life, experiences of positive and negative affect, and occurrences of

mental health symptoms (Fumagalli et al., 2024). If social media use or gaming impacts each of these aspects differently, this could explain why findings to date appear mixed or inconsistent. The first aim of this article, therefore, is to examine how time spent on social media and gaming relates to different aspects of wellbeing, for both boys and girls.

Another debate in the literature is whether any negative impacts of social media use, and more generally online screen time, are practically significant (Pearson, 2025). Meta-analyses generally conclude that the link between social media use and wellbeing is "weak", although a few described the same associations as "substantial" (Valkenburg et al., 2022). One potential reason for these differences in interpretation is that effect size guidelines are inherently "arbitrary", as Cohen (1988) noted when introducing the now widely used correlation coefficient guidelines. Cohen foreshadowed the current debate on how strong a correlation between social media use and wellbeing must be to be considered practically significant, emphasising that this would depend on context. Specifically, Cohen (1988, p. 13) stated:

> "Many effects sought in personality, social, and clinical-psychological research are likely to be small effects as here defined ... Large effects are frequently at issue in such fields as sociology, economics, and experimental and physiological psychology, fields characterized by the study of potent variables or the presence of good experimental control or both."

Some researchers have therefore urged caution against gauging the practical significance of this issue solely based on conventional effect size guidelines used in the field. Instead, they recommend considering more meaningful measures of impact, such as relative risk (Twenge & Hamilton, 2022), which compares the risks of an undesirable outcome occurring between two groups of individuals. Twenge and Hamilton (2022) illustrated the importance of this using an example: while the correlation between smoking and lung cancer was about $r = 0.06$, which is considered "small" by conventional guidelines, the relative risk indicated that smokers were 30 times more likely to develop lung cancer than non-smokers.

The second aim of this article is therefore to examine the relationship between social media use or gaming and adolescent wellbeing, not only using correlation coefficients but also relative risk, to provide a more practical understanding of the potential impacts.

Data

The Programme for International Student Assessment (PISA)[2] provides a valuable source of data for examining the relationship between time spent on social media or gaming and different aspects of wellbeing. PISA is a large-scale international assessment of 15-year-olds, conducted every three years by the Organisation for Economic Co-operation and Development (OECD) since 2000. PISA 2022,

2 The PISA dataset was accessed from the OECD website (https://www.oecd.org/en/data/datasets/pisa-2022-database.html) and analysed following the recommended best practice guidelines (OECD, 2009). The data was fully de-identified (i.e., no student-level identifiers were included), and no results were presented that could rank or evaluate individual schools or students.

the latest survey cycle at the time of writing, is particularly relevant here as it included two optional questionnaires of interest: a wellbeing questionnaire and an Information and Communication Technology (ICT) familiarity questionnaire.

The wellbeing questionnaire provides detailed measures of adolescent wellbeing beyond the routinely asked life satisfaction question, such as frequency of feeling nervous or depressed and the degree to which they agree that they like their appearance just as it is. The ICT questionnaire records, among other things, the availability of digital devices in schools, frequency of using digital devices inside and outside of school, and the frequency of use for specific purposes, such as gaming or browsing social media which is a key piece of information for our analyses.

Therefore, our analyses only focused on countries that administered **both** these optional questionnaires. Table 1 presents the number of sampled students from each of the 10 analysed countries.

Table 1: The number of students sampled in each analysed country

Country	Abbreviated country name	Number of PISA students (unweighted)	Percentage of total sample
Brazil	BRA	10 798	12.3%
Costa Rica	CRI	6 113	6.9%
Hong Kong (China)	HKG	5 907	6.7%
Hungary	HUN	6 198	7.0%
Ireland	IRL	5 569	6.3%
Macao (China)	MAC	4 384	5.0%
Panama	PAN	4 544	5.2%
Saudi Arabia	SAU	6 928	7.9%
Slovenia	SVN	6 721	7.6%
Spain	ESP	30 800	35.0%
Total		**87 962**	**100%**

Although nearly half of the students were from Spain and Brazil, survey weights (known as senate weights in PISA data) were applied to ensure each country contributes roughly[3] equally to the results. Doing so avoids the results being skewed by countries with large numbers of students but instead produces an approximately equal average across all countries.

Social media use and gaming

In the PISA 2022 questionnaire, students were asked how much time they spent browsing "social networks" and playing "video-games (using ... [a] smartphone, a gaming console or an online platform or Apps)" on a typical weekday and

3 The use of senate weights limits each country's contribution to an equivalent of 5 000 observations in the analyses. However, missing data in certain variables can reduce the total weights for some countries, slightly affecting the equality of the average across countries. That said, there are no cases where the results are dominated by a small handful of countries.

weekend day, separately. The exact wording of these PISA questions did not differentiate between usage inside or outside of school, so it is reasonable to assume this reflects students' overall daily use.

Instead of reporting the number of hours, students selected one of six options:

1. No time at all
2. Less than one hour a day
3. Between one and three hours a day
4. More than three hours and up to five hours a day
5. More than five hours and up to seven hours a day
6. More than seven hours a day

In this dataset, 46 per cent of girls spent more than three hours per day browsing social media on weekends and 40 per cent on weekdays. Among boys, the figures were 32 per cent on weekdays and 37 per cent on weekends. Notably, nearly 1 in 4 girls (24 per cent) reported spending more than five hours browsing social media on a typical weekend day.

For gaming, the pattern was reversed between girls and boys. Among boys, 46 per cent reported gaming for more than three hours per day on weekends, dropping to 34 per cent on weekdays. For girls, the figures were much lower: 18 per cent on weekdays and 26 per cent on weekends. Similar to the pattern observed for girls on social media browsing, nearly 1 in 4 boys reported gaming for five hours or more on a weekend day. These observations align with other studies showing that girls tend to spend more time on social media, while boys spend more time gaming.

Wellbeing

The PISA 2022 data included multiple variables, each created by combining student responses to several related questions, designed to measure various aspects of wellbeing. To determine which variables can be considered indicative of adolescent wellbeing, we used the conceptual framework proposed by McLellan and Steward (2015). The authors proposed four aspects of wellbeing for children and adolescents' wellbeing in the school context:

1. Interpersonal wellbeing (e.g., feeling cared for and treated fairly)
2. Life satisfaction
3. Perceived competence (e.g., feeling good about oneself and believing one is doing well)
4. Negative emotion (e.g., feeling worried)

We reviewed the PISA wellbeing questionnaire and identified seven variables that can represent one or more of these aspects. These are life satisfaction, body image perception, general wellbeing on a typical day, sense of belonging at school, feeling of safety, the ability to resist stress, and frequency of experiencing psychosomatic symptoms. The descriptions of these wellbeing variables and their corresponding variable names in the PISA data are presented in Table 2.

Table 2: Wellbeing variables analysed

PISA variable name	Short description	Long description
BELONG	Sense of belonging at school	Students' responses to six statements about feeling like an outsider, making friends easily, belonging at school, feeling awkward or out of place, being liked by peers, and feeling lonely.
BODYIMA	Body image perception	Students' agreement with statements about liking their body, how clothes fit, their appearance, attractiveness, and concerns about weight.
EXPWB	General wellbeing on a typical day	Students' reports of experiencing six positive events the day before the survey: laughing a lot, learning something interesting, having enough energy, being treated with respect, satisfaction with time use, and feeling accomplished.
FEELSAFE	Feeling of safety	Students' reports of feeling of safety in classrooms, other school areas, and on their way to and from school.
LIFESAT	Life satisfaction	Students' satisfaction across 10 areas: health, appearance, learning at school, friends, neighbours, possessions, relationships with parents and teachers, use of time, and school life.
PSYCHSYM	Frequency of experiencing psychosomatic symptoms	Students' ratings of the frequency of nine symptoms related to mind (feeling depressed, irritable, nervous, anxious) and body (headache, stomach pain, back pain, difficulty sleeping, dizziness).
STRESAGR	Ability to resist stress	Students' self-assessment on 10 statements about handling stress, such as getting nervous easily, working under pressure, and worrying about many things.

These wellbeing variables are constructed by the OECD, using Item Response Theory based on student responses to multiple items. Consequently, the values are in logit units, and their interpretation is not straightforward. A score of 0 corresponds to the OECD average for the variable, while scores of +1 and –1 represent one standard deviation above and below this average, respectively. For all but one variable presented in Table 2, higher scores indicate higher levels of wellbeing. The one exception is the frequency of psychosomatic symptoms, for which higher scores indicate lower levels of wellbeing (for more details, see, OECD, 2024[4]).

Not all respondents have data for each wellbeing variable. The amount of missing data varies by variable, mostly because some countries did not administer certain wellbeing questions. For example, respondents from Costa Rica have no data on body image perception (BODYIMA) and life satisfaction (LIFESAT). The variable measuring students' general wellbeing (EXPWB) had the highest level of missing data, mainly from Brazil, Panama, and Saudi Arabia. This is partly because responses from students who answered "no" to the question "Was yesterday a typical day?" were excluded when constructing this variable, as it is intended to reflect wellbeing on a representative day.

4 Specifically, Chapter 19.

Table 3 shows the mean wellbeing scores for boys and girls, along with the standard errors of these estimates. On average, girls report lower wellbeing than boys across all variables. Girls' mean scores fall below the OECD average for variables where higher values indicate greater wellbeing (all except PSYCHSYM) but are above the OECD average in the frequency of psychosomatic symptoms (PSYCHSYM).

Table 3: Summary statistics for wellbeing scores by gender

Wellbeing variable	Mean score			Standard error around the mean estimate	
	Girls	Boys	Difference (girls-boys)	Girls	Boys
BELONG	-0.15	-0.02	-0.13	0.006	0.006
BODYIMA	-0.25	0.01	-0.26	0.009	0.007
EXPWB	-0.08	0.13	-0.21	0.008	0.007
FEELSAFE	-0.22	0.03	-0.25	0.008	0.007
LIFESAT	-0.12	0.05	-0.17	0.008	0.009
PSYCHSYM	0.24	-0.35	0.59	0.006	0.007
STRESAGR	-0.33	0.26	-0.59	0.007	0.007

Methods

Descriptive statistics

We began the analysis by examining the mean wellbeing scores by students' time spent on social media and gaming, separately for girls and boys. This is a straightforward way to assess whether more time spent on these activities is associated with lower wellbeing. We then calculated Spearman's correlation coefficients to quantify the strength of these relationships.[5] For comparison, we also calculated the correlation coefficients between frequency of being bullied[6] and wellbeing, as it is reasonable to expect that wellbeing would be affected by whether students are bullied. Doing so helps us interpret the effect sizes for time spent on digital activities in relation to a factor more widely recognised as having a substantial impact on wellbeing. This supports the second aim of this article to provide a clearer understanding of the practical significance of these associations.

For this and all other analyses in this article, senate weights were applied to ensure roughly equal contributions from each country, and replicate weights (as provided in the PISA dataset but readjusted to ensure each country is equally

5 Spearman's correlation was used instead of the Pearson's correlation because the time use variable is ordinal (e.g., zero hours, less than one hour, ..., more than seven hours). The calculations were performed in R using the *intsvy* (Caro & Biecek, 2017) and *wCorr* (Bailey et al., 2023) R packages.

6 This data was collected by PISA and readily available in the dataset.

weighted)[7] were used to account for sampling variance when calculating standard errors.

Relative risk

To better understand the relationships between time spent browsing social media and gaming and the different aspects of wellbeing in practical terms, we calculated the relative risk of being in the bottom 25 per cent[8] of wellbeing scores within each gender group, comparing adolescents (i) who browsed social media for more than three hours daily with (ii) those who browsed for three hours or less. These relative risks were calculated separately for girls and boys and for each wellbeing variable. We did the same calculation for gaming, comparing adolescents who gamed for more than three hours daily with those who gamed for fewer hours.

There is no consensus on the optimal amount of time for gaming or social media use, so any threshold dividing students into two groups for relative risk calculations is necessarily arbitrary. Previous studies have used different threshold values: for example, Twenge et al. (2020) compared those spending two or more hours on the internet with those spending less than two, and also contrasted very heavy users (over seven hours) with light users (30 minutes to one hour). Similarly, Twenge et al. (2018) examined relative risks between three or more hours of electronic device use versus less than three.

Our threshold choice is guided by Zablotsky (2025), who defined high "non-schoolwork screen time" as four or more hours per day and found that adolescents in this group were more likely to experience adverse health outcomes such as depression and anxiety. Since the definition of "screen time" in that study likely included time spent beyond social media browsing and gaming, we considered the "more than three hours and up to five hours" category in our data to be the most appropriate threshold here.

A relative risk of 1 means both groups are equally likely to be in the bottom 25 per cent of wellbeing scores. Values above 1 indicate a higher likelihood for those spending more than three hours, while values below 1 indicate a lower likelihood compared to those spending less time. For easier interpretation, the score for psychosomatic symptoms (PSYCHSYM) was reversed for this analysis, so a relative risk greater than 1 indicates a higher likelihood of being in the bottom quartile group that experiences these symptoms most frequently.

Regression analysis

To assess the relationship between time spent browsing social media or gaming and wellbeing, while controlling for student background characteristics, we conducted multivariate regression analyses estimated using Ordinary Least Squares. For each wellbeing outcome, two models were fitted:

7 Since the Balanced Repeated Replication (BRR) weights in the PISA dataset assign larger total weights to countries with bigger populations, we scaled the replicate weights so that each country had the same total weight within each set of replicate weights. This was achieved by applying a constant factor to each observation within a country, bringing its total weight to 5 000, while maintaining relative differences between observations. These adjustments were applied to all sets of replicate weights and used in all analyses.

8 This was calculated using weighted percentile using the wtd.quantile() R function from the *Hmisc* package (Harrell Jr, 2024).

1. **Model 1**: Includes all control variables and the time spent browsing social media or gaming on weekends and weekdays, separately.

2. **Model 2**: Builds on Model 1 by adding four interaction terms: the interaction between time spent browsing social media or gaming during weekends and gender, and the interaction between time spent browsing social media or gaming during weekdays and gender.

Model 1 estimates the association between browsing social media and wellbeing, and between gaming and wellbeing for all students, while controlling for students' background characteristics. Model 2 explores whether these associations differ by gender through interaction effects.

The models included the following control variables:[9]

- **student characteristics**: age, grade, gender, academic performance as approximated by student math performance in PISA, whether they have recently skipped classes;
- **socioeconomic and family environment**: economic, social, and cultural status index (ESCS), family support, frequency of doing paid work and household duties;
- **interpersonal relationships**: days spent with friends, number of close friends, relationships with teachers, frequency of being bullied;
- **physical health indicators**: exercise frequency, body mass index; and
- **school background**: student–teacher ratio, computer–student ratio, school type, negative school climate.

Our regression estimates are likely to be conservative due to potential overcontrolling. To estimate a direct effect closer to causal, that is, how wellbeing changes with increases in time spent browsing social media or gaming, the regression models controlled for factors related to both wellbeing and time spent browsing social media or gaming, such as exercise frequency, days spent with friends, and frequency of doing paid work (as shown in the preceding list). However, if time spent on digital activities replaces in-person activities beneficial for wellbeing (time displacement theory, see Moy et al., 1999; Putnam, 2000), then our results may have underestimated the total impact of digital use on wellbeing as indirect effects are not fully captured here.

For all regression analyses, we applied a stricter significance level than the conventional 5 per cent due to the large number of hypotheses tested. As such, we only considered an effect to be statistically significant when $p \leq 0.01$. The mathematical specification for both models, including expanded descriptions of the control variables, can be found in Table A1 in the Appendix.

9 Data for these control variables was obtained from both the student and school PISA datasets. The student dataset includes responses from the student questionnaire as well as other questionnaires, such as the wellbeing and ICT familiarity questionnaires. The school dataset consists of responses provided by school principals through the school questionnaire.

Results

Descriptive statistics

Figure 1 shows the mean wellbeing scores by students' time spent browsing social media or gaming, presented separately for girls and boys. Solid dots represent means based on weekend usage, while hollow dots represent weekday usage. However, given that means for weekday and weekend usage were very similar, these dots largely overlap in the graphs.

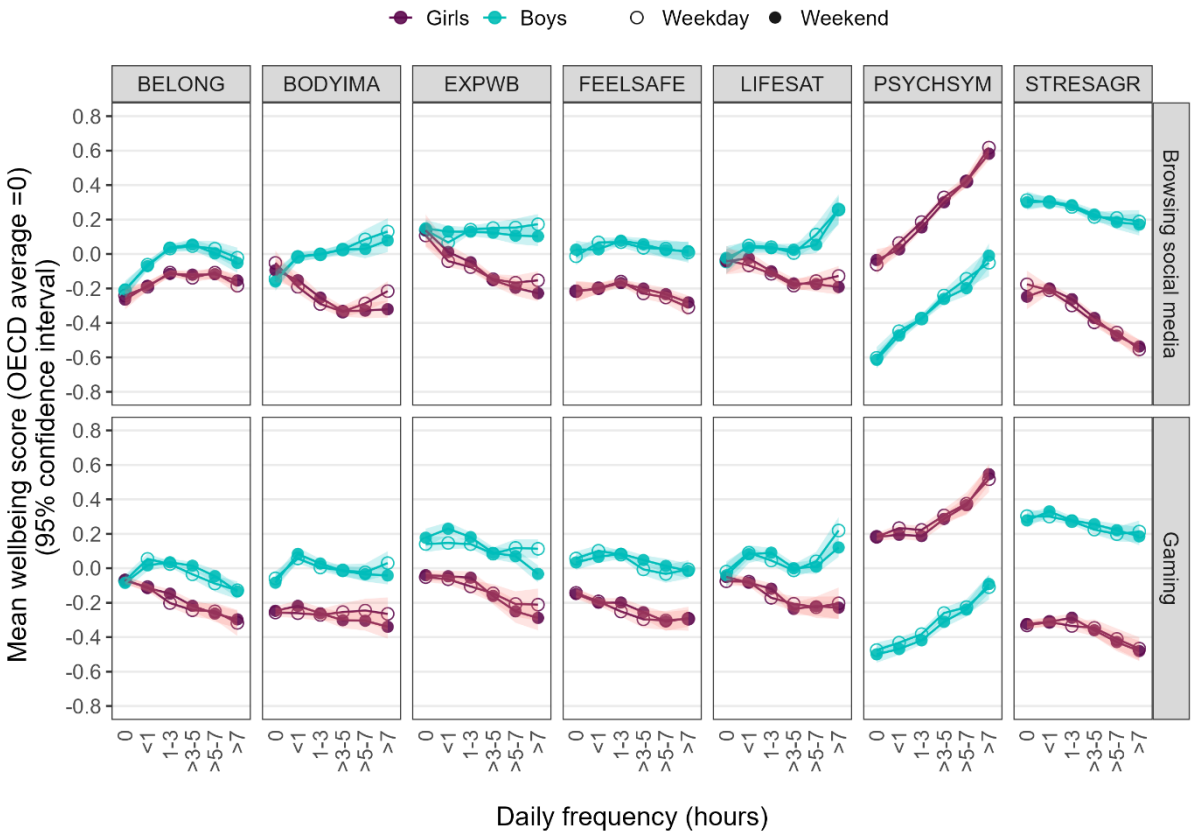

Figure 1: Mean wellbeing scores (with 95 per cent confidence interval) by time spent on social media or gaming during a typical weekday and weekend day, for girls and boys

The first observation from the figure is that relationships between time spent on social media or gaming and mean wellbeing scores varied across wellbeing variables. The wellbeing aspect that was most visibly negatively linked to time spent on these uses was the frequency of experiencing psychosomatic symptoms (PSYCHSYM), that is, those who spent longer browsing social media or gaming experienced psychosomatic symptoms more often. Students who spent over seven hours browsing social media or gaming reported PSYCHSYM scores that were, on average, about 0.5 logits higher than those spending less than one hour. This magnitude of difference was observed for both girls and boys. In contrast, a positive association to wellbeing, though smaller, was observed for sense of belonging at school (BELONG). For example, girls who browsed for more than one hour daily, on average, reported higher BELONG than girls who browsed less or not at all.

The second observation is that there were several cases where the relationship differed between girls and boys. Girls who spent more time browsing social media, on average, reported lower body image perception, lower general wellbeing, lower stress resistance, and, to a smaller extent, lower life satisfaction. In contrast, these mean wellbeing scores remained relatively similar among boys regardless of their usage or only slightly declined as time spent browsing social media increased. Boys who spent more than seven hours browsing social media had a noticeably higher average life satisfaction score than those who spent fewer hours browsing, but the reason for this is unclear. For the relationships between time spent gaming and wellbeing, there were fewer wellbeing variables that showed a clear gender difference.

Figure 2 shows the correlation coefficients for these relationships based on time spent during weekends, with correlations between frequency of being bullied and wellbeing included for comparison. Results based on weekday usage are very similar and, hence, are not presented here.

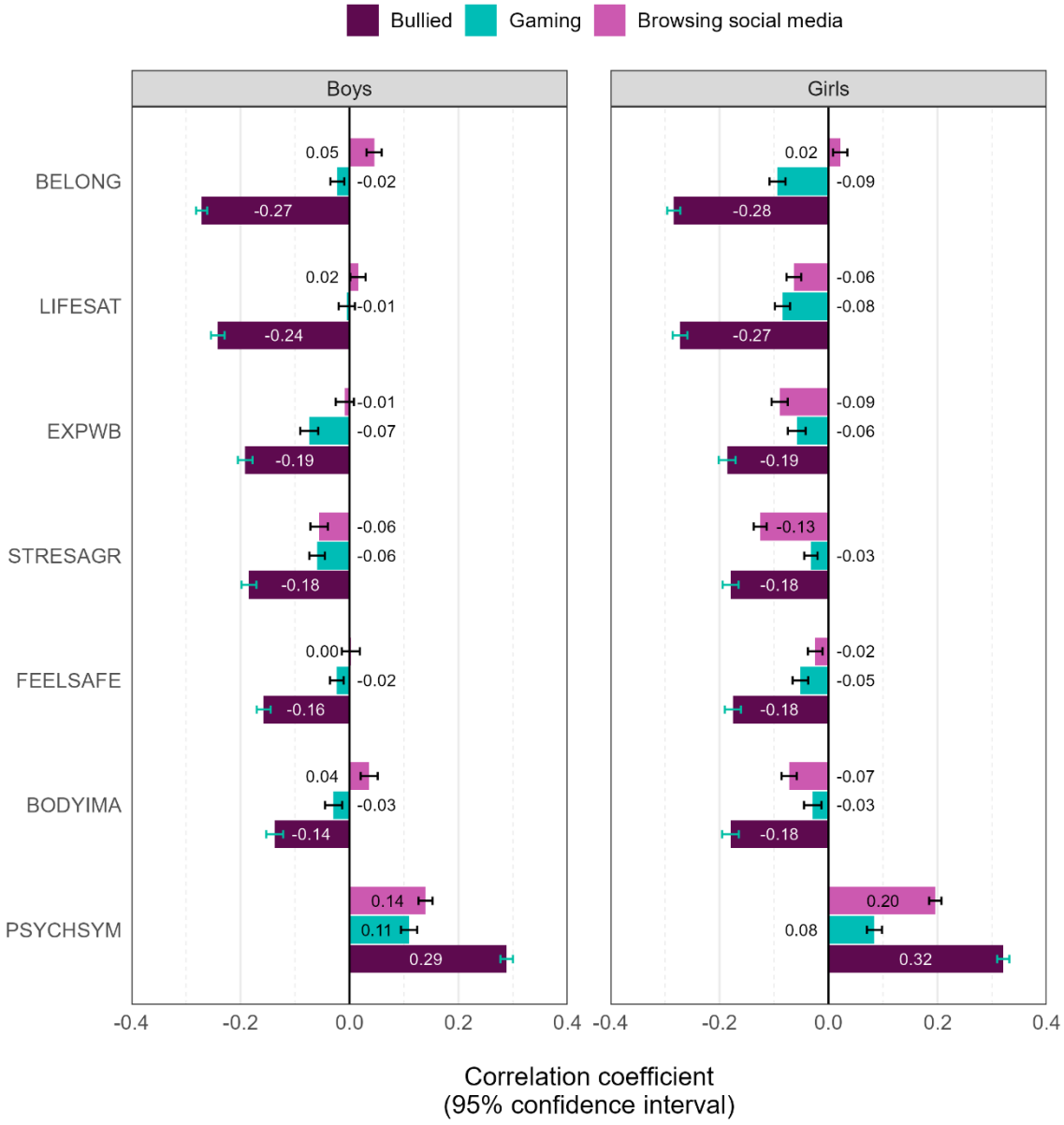

Figure 2: Spearman's rank correlation coefficients between wellbeing scores and (i) time spent on social media or gaming on a typical weekend day, and (ii) frequency of being bullied

As shown in the figure, for girls, the strongest correlation was between time spent browsing social media and frequency of psychosomatic symptoms (r = 0.20), which was about 63 per cent of the correlation between frequency of being bullied and psychosomatic symptoms (r = 0.32). The second strongest correlation was between time spent browsing social media and the ability to resist stress (r = -0.13), about 70 per cent of the correlation between being bullied and stress resistance (r = -0.18). For gaming, the correlations were smaller, ranging (in absolute terms) from 0.03 to 0.09 depending on the wellbeing variable.

For boys, the two strongest correlations were between psychosomatic symptoms and browsing social media (r = 0.14) and between psychosomatic symptoms and gaming (r = 0.11), corresponding to about 48 per cent and 38 per cent of the correlation between psychosomatic symptoms and being bullied, respectively.

Relative risk

When comparing the risk of having a wellbeing score in the lowest quartile (i.e., bottom 25 per cent) between students who spent more than three hours and those who spent less time, a similar pattern emerges, as shown in Figure 3: (i) the level of relative risk varies by the wellbeing outcome considered, and (ii) some gender differences were evident.

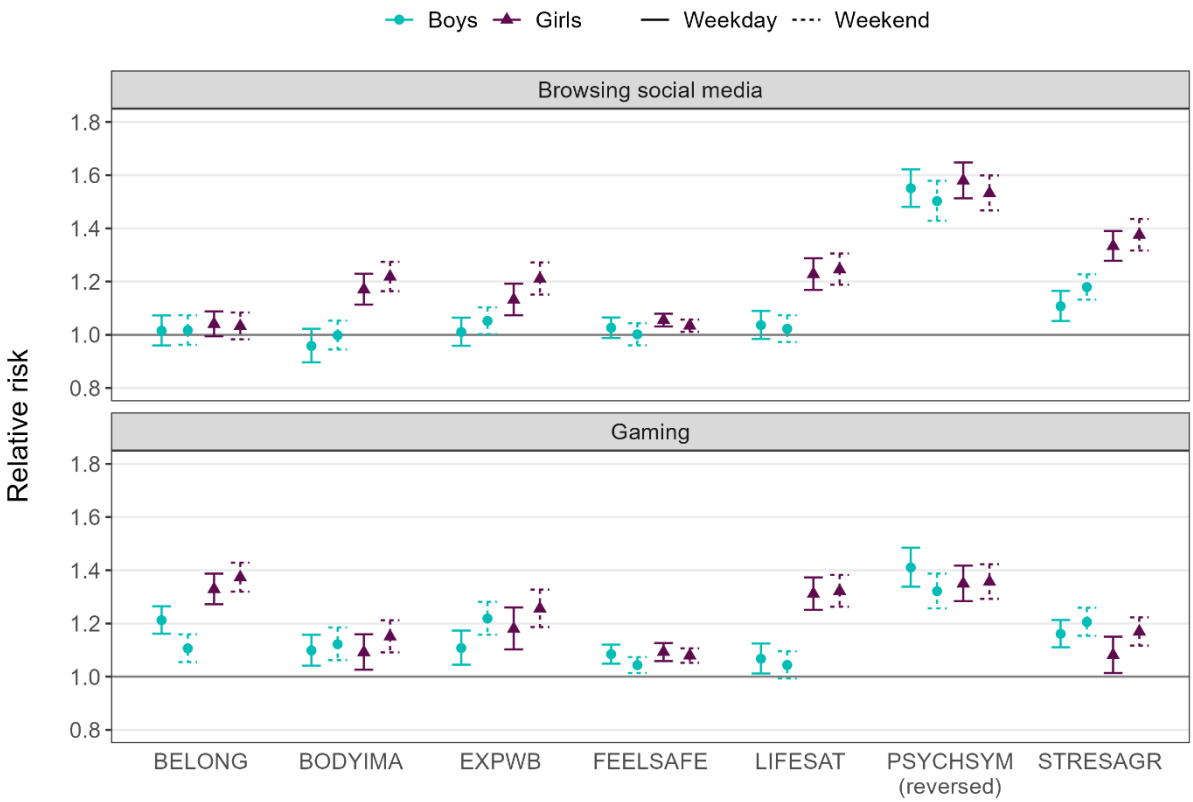

Figure 3: Relative risk of being in the bottom 25 per cent of wellbeing scores (within each gender) for students spending more than three hours daily browsing social media or gaming compared to those spending less

Girls who browsed social media for more than three hours per day were about 20 per cent more likely (i.e., relative risk value of around 1.2) to have body image perception, general wellbeing, and life satisfaction scores in the lowest quartile compared to girls who browsed less. They were also about 30 per cent more likely to fall into the lowest quartile for stress resistance and at least 50 per cent more likely to do so for psychosomatic symptoms. In contrast, this pattern was not observed among boys. Boys were generally equally likely to be in the lowest quartile for all wellbeing variables, except for psychosomatic symptoms. For psychosomatic symptoms, boys who browsed for more than three hours were about 50 per cent more likely to be in the lowest quartile than boys who browsed less or not at all.

Relative risks for gaming were generally lower than for browsing social media. Among girls, the highest relative risks for gaming were observed for sense of belonging, life satisfaction, and psychosomatic symptoms. Girls who gamed for more than three hours were about 30 per cent more likely to be in the lowest quartile for these wellbeing variables compared to girls who gamed less or not at all. For boys, the highest relative risk was again for psychosomatic symptoms, where those who gamed for more than three hours were at least 30 per cent more likely to be in the lowest quartile than those who gamed less.

Regression analysis

Table 4 presents the key estimated regression coefficients from Model 1, which shows the associations between time spent gaming or browsing social media (separately for weekend and weekday) and each wellbeing outcome for all students overall. Similarly, Table 5 presents results from Model 2, which shows whether these associations vary by gender. Full regression outputs from both models are available in the Supplementary Materials.

Overall, even after controlling for observable characteristics and background variables, the patterns observed earlier remained evident.

Firstly, as shown in Table 4, greater time spent browsing social media was statistically significantly associated with a higher frequency of psychosomatic symptoms, with a difference of approximately 0.30 logits[10] between those browsing more than seven hours daily and non-browsers. Time spent gaming, specifically during weekends, was also associated with higher frequency of psychosomatic symptoms but to a much lesser extent compared to browsing social media.

Secondly, increased time spent browsing social media was associated with lower stress resistance. For instance, compared to non-browsers, those who browsed more than seven hours per day had stress resistance scores that were 0.15–0.18 logits lower, depending on whether usage occurred on a weekday or weekend day. These relationships did not differ significantly by gender (see the interaction terms in Table 5), and the same negative association was not found in time spent gaming.

10 This was calculated by multiplying the estimated coefficient for browsing social media by the difference in numeric category between those who spent seven or more hours per day (coded as six) and those who spent no time on this activity (coded as one), i.e., 0.06 logits × (6 – 1) = 0.06 × 5 = 0.30 logits.

Thirdly, positive associations were observed primarily between a sense of belonging at school and time spent browsing social media. For instance, students who browsed for more than seven hours daily were estimated to have belonging scores that were 0.11 logits higher than those of non-browsers, controlling for their background characteristics (or 0.09 logits higher compared to those browsing for less than one hour). These magnitudes were noticeably smaller than the negative associations observed. This relationship did not differ significantly between girls and boys (see the interaction terms in Table 5).

Fourthly, the associations between gaming and wellbeing are different to the associations between browsing social media and wellbeing in two cases: increased time spent gaming on weekdays is associated with lower sense of belonging at school and positively associated with body image perception. These relationships did not differ significantly between girls and boys (see the interaction terms in Table 5).

Lastly, results from Table 5 indicate that the associations between time spent gaming and browsing social media and wellbeing differed statistically between girls and boys for body image perception, general wellbeing, and life satisfaction. These gender differences are indicated by the interaction terms (e.g., "Browsing SM (Weekday) * Boys"). To ease the interpretation of these differences, Figure 4 illustrates the predicted wellbeing scores for variables with statistically significant gender interaction terms.[11] As shown in the figure, more time spent browsing social media on weekends was negatively associated with body image perception, for girls but not boys. Furthermore, while general wellbeing is expected to be lower for those who spent more time browsing social media on weekends, the decline in general wellbeing as time spent increases was estimated to be much larger for girls than for boys. For life satisfaction, neither browsing social media nor gaming showed a statistically significant overall association (see Table 4). However, the association differed by gender, where boys' life satisfaction tended to be higher with more time spent browsing social media or gaming, whereas girls' life satisfaction tended to decline instead (Figure 4).

11 Figure 4 presents the predicted wellbeing scores for students with a specific set of background characteristics. However, the overall relationships between time spent gaming or browsing social media and wellbeing are expected to be the same for students with different background profiles. The only difference lies in the level of the predicted scores, which may be higher or lower depending on the students' characteristics.

Table 4: Key estimated coefficients from regression Model 1

Variable	Dependent variable						
	BELONG	BODYIMA	EXPWB	FEELSAFE	LIFESAT	PSYCHSYM	STRESAGR
(Intercept)	-0.309 (0.238)	-1.148 (0.326) **	0.540 (0.321)	-0.956 (0.305) *	0.226 (0.310)	-0.418 (0.275)	-1.067 (0.294) **
Boys	0.107 (0.009) **	0.219 (0.012) **	0.169 (0.012) **	0.229 (0.011) **	0.127 (0.014) **	-0.542 (0.011) **	0.495 (0.011) **
Browsing SM (Weekday)	0.022 (0.005) **	0.008 (0.007)	-0.000 (0.007)	0.011 (0.006)	-0.004 (0.006)	0.060 (0.006) **	-0.029 (0.005) **
Browsing SM (Weekend)	0.022 (0.005) **	-0.017 (0.006) *	-0.033 (0.006) **	0.000 (0.006)	0.000 (0.006)	0.056 (0.005) **	-0.035 (0.005) **
Gaming (Weekday)	-0.020 (0.005) **	0.020 (0.005) **	0.004 (0.006)	-0.013 (0.006)	0.010 (0.006)	-0.003 (0.005)	0.012 (0.006)
Gaming (Weekend)	-0.009 (0.005)	0.001 (0.004)	-0.023 (0.005) **	-0.003 (0.005)	0.009 (0.005)	0.024 (0.005) **	-0.001 (0.005)
R-squared	0.199 (0.004)	0.149 (0.004)	0.180 (0.004)	0.124 (0.004)	0.224 (0.004)	0.214 (0.003)	0.159 (0.004)
Weighted count	38 895	32 933	29 260	34 973	34 874	38 799	32 353

Notes: Digital usage variables were coded as numeric values from 1 (0 hours), 2 (< 1 hour), ... , to 6 (> 7 hours). "SM" denotes social media. Regression coefficients significant at p values ≤ 0.001 are indicated by **, and * for those at $p \leq 0.01$. Standard errors (in parentheses) were estimated using Balanced Repeated Replication weights, which were adjusted to ensure equal country representation. The linear model included country fixed effects and a set of student, family, and school characteristics. The full outputs are available in the Supplementary Materials.

Table 5: Key estimated coefficients from regression Model 2

Variable	Dependent variable						
	BELONG	BODYIMA	EXPWB	FEELSAFE	LIFESAT	PSYCHSYM	STRESAGR
(Intercept)	-0.269 (0.235)	-1.045 (0.324) **	0.626 (0.321)	-0.937 (0.305) *	0.342 (0.305)	-0.423 (0.274)	-0.992 (0.299) **
Boys	0.035 (0.025)	0.026 (0.033)	0.022 (0.030)	0.197 (0.029) **	-0.088 (0.034) *	-0.541 (0.029) **	0.377 (0.032) **
Browsing SM (Weekday)	0.013 (0.007)	0.005 (0.009)	-0.007 (0.009)	0.010 (0.008)	0.000 (0.008)	0.061 (0.008) **	-0.040 (0.008) **
Browsing SM (Weekend)	0.022 (0.007) **	-0.042 (0.009) **	-0.050 (0.009) **	-0.005 (0.007)	-0.018 (0.007)	0.063 (0.007) **	-0.048 (0.006) **
Gaming (Weekday)	-0.024 (0.007) **	0.022 (0.008) *	-0.011 (0.008)	-0.015 (0.008)	-0.009 (0.008)	-0.005 (0.007)	0.020 (0.008) *
Gaming (Weekend)	-0.008 (0.006)	-0.001 (0.008)	-0.009 (0.008)	0.001 (0.007)	0.005 (0.006)	0.018 (0.006) *	-0.002 (0.007)
Browsing SM (Weekday) * Boys	0.020 (0.011)	0.008 (0.012)	0.014 (0.013)	0.003 (0.012)	-0.009 (0.013)	-0.003 (0.012)	0.023 (0.012)
Browsing SM (Weekend) * Boys	-0.001 (0.010)	0.055 (0.012) **	0.037 (0.014) *	0.012 (0.012)	0.034 (0.011) *	-0.016 (0.009)	0.028 (0.011)
Gaming (Weekday) * Boys	0.006 (0.009)	-0.007 (0.012)	0.024 (0.012)	0.003 (0.011)	0.033 (0.011) *	0.006 (0.010)	-0.019 (0.011)
Gaming (Weekend) * Boys	-0.000 (0.009)	0.002 (0.012)	-0.028 (0.012)	-0.009 (0.010)	0.011 (0.010)	0.014 (0.009)	0.000 (0.011)
R-squared	0.199 (0.004)	0.151 (0.004)	0.181 (0.004)	0.124 (0.004)	0.226 (0.004)	0.214 (0.003)	0.160 (0.003)
Weighted count	38 895	32 933	29 260	34 973	34 874	38 799	32 353

Notes: Digital usage variables were coded as numeric values from 1 (0 hours), 2 (< 1 hour), ... , to 6 (> 7 hours). "SM" denotes social media. Regression coefficients significant at p values ≤ 0.001 are indicated by **, and * for those at $p \leq 0.01$ Standard errors (in parentheses) were estimated using Balanced Repeated Replication weights, which were adjusted to ensure equal country representation. The linear model included country fixed effects and a set of student, family, and school characteristics. The full outputs are available in the Supplementary Materials.

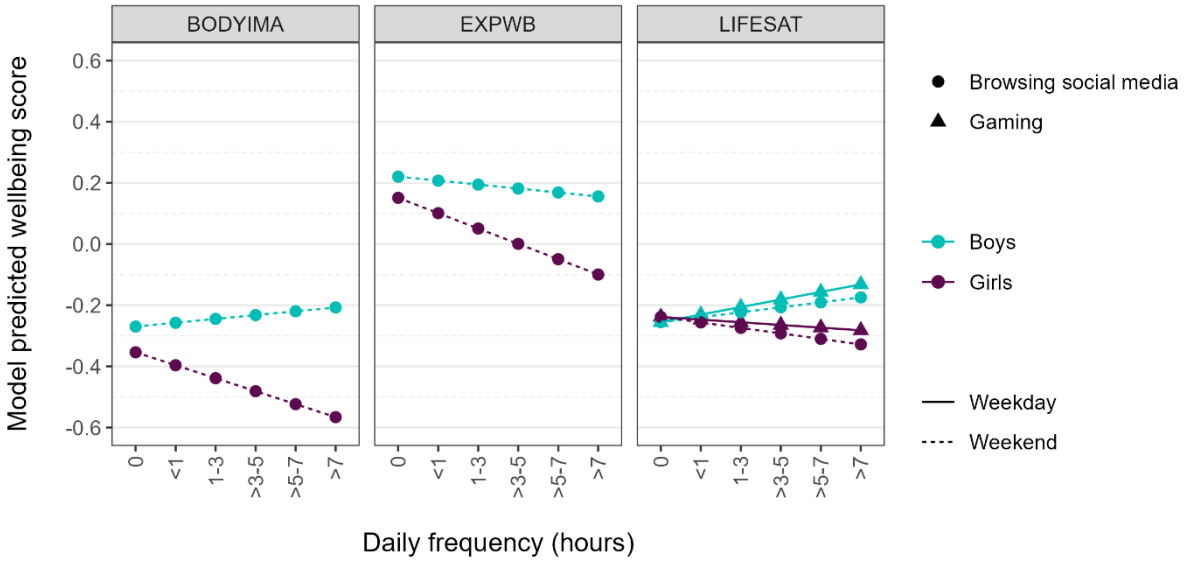

Figure 4: Predicted wellbeing scores by gender based on Model 2. Predictions were for 15-year-old students from group 5 of Economic, Social and Cultural Status, with medium levels of family support, number of close friends, quality of relationship with teachers, frequency of bullying, and body mass index. Students were in grade 10, not skipping classes two weeks prior to the PISA test, not working outside of school, spending two days per week with friends after school, not exercising, from Ireland, studying in a public school with medium student–teacher ratio, computer–student ratio, negative school climate, and a plausible math score of 450.

Discussion and conclusion

Using data from PISA 2022, this article first aimed to examine the relationship between time spent browsing social media or gaming and various aspects of adolescent wellbeing. The intention was to shed light on the mixed findings in the existing literature by exploring whether differing definitions and conceptualisations of wellbeing could have contributed to this heterogeneity. The second aim was to investigate the practical significance of these relationships by going beyond simple correlation coefficients, which are commonly used in research based on cross-sectional data. This aim was motivated by the ongoing debate about whether the correlations found in research of this kind hold any practical significance. This has implications for policy decisions on whether and which measures are likely to be effective in protecting adolescent wellbeing, which is on a declining trajectory.

Across seven aspects of adolescent wellbeing, we found that both the direction and strength of associations to social media browsing and gaming varied depending on the specific wellbeing aspect considered. The biggest negative association observed was between time spent browsing social media and the frequency of psychosomatic symptoms, such as headaches and anxiety. Similar associations were observed for both girls and boys, and the patterns held even after controlling for students' background characteristics. For example, the correlation between social media browsing and psychosomatic symptoms was about 63 per cent of the correlation between being bullied and psychosomatic symptoms for girls, and about 48 per cent for boys. Adolescents who browsed for

more than three hours a day were about 50 per cent more likely to fall within the top quartile experiencing the most frequent psychosomatic symptoms, compared to their peers of the same gender. Time spent gaming also had the same link with psychosomatic symptoms, but to a lesser extent. A consistent, though weaker, negative association was also found for stress resistance. For example, girls who spent three or more hours browsing social media were about 30 per cent more likely to be in the bottom 25 per cent for stress resistance than girls who spent less time.

Evidence of positive associations was mainly observed for students' sense of belonging at school. Those who engaged in light browsing (one to three hours per day) reported a higher sense of belonging compared to peers who browsed less or not at all. However, sense of belonging declined as browsing time exceeded three hours per day. This may explain why, when examining relative risks, students who spent more than three hours browsing were not significantly more likely to have low belonging scores (in the lowest quartile) than those who browsed less. These positive associations were more noticeable after controlling for student background characteristics. Nevertheless, the magnitude of the negative associations with psychosomatic symptoms or stress resistance outweighed the positive association with sense of belonging.

The variations in the type and strength of associations highlight the need to clearly define specific wellbeing outcomes in future research, particularly for meta-analyses that synthesise findings across diverse studies. Moreover, users of research should also understand that not all aspects of adolescent wellbeing are equally vulnerable to the effects of social media browsing and gaming.

Furthermore, we also found that the associations between time spent gaming or browsing social media and wellbeing differed between girls and boys, particularly for body image perception, general wellbeing, and life satisfaction. Girls who browsed social media for more than three hours per day were about 20 per cent more likely to score in the lowest quartile for these measures compared to girls who browsed less. Similarly, girls who gamed for more than three hours were 30 per cent more likely to have life satisfaction scores in the lowest quartile. These patterns were not observed among boys, and the gender differences remained even after controlling for student background characteristics. This finding highlights the necessity for future studies to examine gender-specific associations or impact and the need for policymakers to develop targeted measures that address the challenges faced by girls in the digital world.

Limitations

As with any study based on cross-sectional and non-experimental data, the associations reported in this article cannot be interpreted as causal. Firstly, the analysis relies on self-reported measures of wellbeing and time spent browsing social media and gaming, which are susceptible to inaccuracies. Secondly, it is possible that time spent on these activities and wellbeing influence each other simultaneously. Consequently, for example, the observed positive association between sense of belonging at school and social media browsing may partly reflect that students with more friends are more likely to browse social media.

Furthermore, as noted in the Methods section, our regression estimates may have underestimated the total impact of time spent on browsing social media or gaming on wellbeing because we controlled for factors such as exercise frequency and time spent with friends. As a result, our model does not capture any indirect effects on adolescents' wellbeing that may occur through the displacement of beneficial in-person activities.

Lastly, while we found a negative association between time spent gaming and browsing social media and adolescents' experience of psychosomatic symptoms, we were unable to determine whether these associations were driven primarily by physical symptoms which are potentially linked to sedentary behaviour, or by mental symptoms, which may relate to the experiences of online interactions and content exposure. This limitation arises because the wellbeing measure used here was a composite that combines both physical and mental aspects. Future research should examine these dimensions separately, as doing so could help identify which interventions are most likely to be effective in addressing adolescent experience of psychosomatic symptoms.

References

Bailey, P., Sikali, E., Emad, A., & Buehler, E. (2023). wCorr: Weighted correlations (R package version 1.9.8) [Computer software].

Boniel-Nissim, M., Marino, C., Galeotti, T., Blinka, L., Ozoliņa, K., Craig, W., Lahti, H., Wong, S. L., Brown, J., Wilson, M., Inchley, J., & van den Eijnde, R. (2024). A focus on adolescent social media use and gaming in Europe, central Asia and Canada. Health Behaviour in School-aged Children international report from the 2021/2022 survey (Vol. 6). WHO Regional Office for Europe.

Caro, D. H., & Biecek, P. (2017). intsvy: An R package for analyzing international large-scale assessment data. *Journal of Statistical Software, 81*(7), 1–44.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Fumagalli, E., Shrum, L. J., & Lowrey, T. M. (2024). The effects of social media consumption on adolescent psychological wellbeing. *Journal of the Association for Consumer Research, 9*(2), 119–130.

Harrell Jr, F. E. (2024). *Hmisc: Harrell miscellaneous*. (R package version 5.2-1) [Computer software].

Gutman, L. M., & Vorhaus, J. (2012). *The impact of pupil behaviour and wellbeing on educational outcomes*. The Childhood Wellbeing Research Centre.

Jerrim, J., Lopez-Agudo, L. A., & Marcenaro-Gutierrez, O. D. (2022). The impact of test language on PISA scores. New evidence from Wales. *British Educational Research Journal, 48*(3), 420–445.

Kaya, M., & Erdem, C. (2021). Students' well-being and academic achievement: A meta-analysis study. *Child Indicators Research, 14*(5), 1743–1767.

Keaten, J. (2025, November 8). Another country agrees to ban social media for children under 15. *Independent*.

Marquez, J., Taylor, L., Boyle, L., Zhou, W., & De Neve, J.-E. (2024). Child and adolescent wellbeing: Global trends, challenges and opportunities. In J. F. Helliwell, R. Layard, J. D. Sachs, J.-E. De Neve, L. B. Aknin, & S. Wang (Eds.), *World Happiness Report 2024* (pp. 61–102). Wellbeing Research Centre, University of Oxford.

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

Moy, P., Scheufele, D. A., & Holbert, R. L. (1999). Television use and social capital: Testing Putnam's time displacement hypothesis. *Mass Communication and Society, 2*(1–2), 27–45.

OECD. (2009). PISA data analysis manual: SAS (2nd ed.). OECD Publishing.

OECD. (2024). *PISA 2022 technical report*. OECD Publishing.

Orben, A., Przybylski, A. K., Blakemore, S.-J., & Kievit, R. A. (2022). Windows of developmental sensitivity to social media. *Nature Communications, 13*(1), 1649.

Pearson, H. (2025, April 2). Do smartphones and social media really harm teens' mental health? *Nature*.

Putnam, R. (2000). *Bowling alone: The collapse and revival of American community.* Simon and Schuster.

Radesky, J., Weeks, H. M., Schaller, A., Robb, M., Mann, S., & Lenhart, A. (2023). *Constant companion: A week in the life of a young person's smartphone use*. Common Sense Media.

Rothwell, J. (2023, October 13). Teens spend average of 4.8 hours on social media per day. *Gallup*.

Svensson, R., Johnson, B., & Olsson, A. (2022). Does gender matter? The association between different digital media activities and adolescent well-being. *BMC Public Health, 22,* 273.

The Children's Society. (2025). *The Good Childhood Report 2025.*

Twenge, J. M., Haidt, J., Joiner, T. E., & Campbell, W. K. (2020). Underestimating digital media harm. *Nature Human Behaviour, 4,* 346–348.

Twenge, J. M., & Hamilton, J. L. (2022). Linear correlation is insufficient as the sole measure of associations: The case of technology use and mental health. *Acta Psychologica, 229,* 103696.

Twenge, J. M., Joiner, T. E., Rogers, M. L., & Martin, G. N. (2018). Increases in depressive symptoms, suicide-related outcomes, and suicide rates among U.S. adolescents after 2010 and links to increased new media screen time. *Clinical Psychological Science, 6*(1), 3–17.

Valkenburg, P. M., Meier, A., & Beyens, I. (2022). Social media use and its impact on adolescent mental health: An umbrella review of the evidence. *Current Opinion in Psychology, 44,* 58–68.

World Health Organization. (2025, August 29). *Depressive disorder (depression).*

Zablotsky, B., Ng, A. E., Black, L. I., Haile, G., Bose, J., Jones, J. R., & Blumberg, S. J. (2025). Associations between screen time use and health outcomes among US teenagers. *Preventing Chronic Disease, 22,* E38.

Appendix

Specification of the regression models

The mathematical specification of the regression Model 2 is as follows:

$$w_{isc} = \beta_0 + \boldsymbol{\beta_1}(\boldsymbol{D_{isc}}) + \boldsymbol{\beta_2}(\boldsymbol{D_{isc}} * G_{isc}) + \boldsymbol{\beta_3 X_{isc}} + \beta_4 G_{isc} + \boldsymbol{\theta_c} + \varepsilon_{isc}$$

where i denotes student, s indexes school and c indexes country. The variable w_{isc} represents the wellbeing of student i from school s in country c.

$\boldsymbol{D_{isc}}$ is the set of variables indicating time spent browsing social media or gaming (henceforth, digital usage variables), and $\boldsymbol{\beta_1}$ represents the effects of digital usage on students' wellbeing, averaged across countries.

G_{isc} is the gender variable (1 = boys, 0 = girls), and $\boldsymbol{D_{isc}} * G_{isc}$ represents the gender interaction terms. The coefficients of interest here are $\boldsymbol{\beta_2}$ which capture whether the association between time spent browsing social media or gaming varies by gender.

$\boldsymbol{X_{isc}}$ denotes the vector of control variables (see Table A1 for full descriptions and coding; the gender and country terms were only included once). The term $\boldsymbol{\theta_c}$ indicates the country fixed effect to account for unobserved country characteristics (e.g., country's wealth). Lastly, the term ε_{isc} is the error term representing factors not captured in the model that affect w_{isc}.

Model 1 has a similar specification to Model 2, but it excludes the gender interaction term $\boldsymbol{\beta_2}(\boldsymbol{D_{isc}} * G_{isc})$. The coefficients of interest in Model 1 are therefore $\boldsymbol{\beta_1}$ which indicate whether the digital usage is a significant predictor of wellbeing.

Although we did not use a multilevel model, which is common in educational research, our approach accounts for some of the hierarchical structure of the data through the use of Balanced Repeated Replication (BRR) weights, country fixed effects, and school-level variables. As noted by OECD (2009), "while simple linear regression models do not recognise hierarchical structure of data, it is possible to account for some hierarchical aspects of the PISA data [...] in the linear regression by using BRR weights [...] These models can adjust for clustering of students within schools and other aspects of survey design" (p. 229). This approach was also used in prior research (e.g., Jerrim et al., 2022).

All regression analyses were conducted using the *intsvy* R package (Caro & Biecek, 2017).

Description of control variables

Table A1: Description of control variables included in the regression models

Regression variable name	PISA variable name	Description	Variable values*
BRA	CNT	Country name – Brazil	1 = Brazil, 0 if otherwise
CRI	CNT	Country name – Costa Rica	1 = Costa Rica, 0 if otherwise
HKG	CNT	Country name – Hong Kong	1 = Hong Kong, 0 if otherwise
HUN	CNT	Country name – Hungary	1 = Hungary, 0 if otherwise
IRL	CNT	Country name – Ireland	1 = Ireland, 0 if otherwise
MAC	CNT	Country name – Macao	1 = Macao, 0 if otherwise
PAN	CNT	Country name – Panama	1 = Panama, 0 if otherwise
SAU	CNT	Country name – Saudi Arabia	1 = Saudi Arabia, 0 if otherwise
SVN	CNT	Country name – Slovenia	1 = Slovenia, 0 if otherwise
ESP	CNT	Country name – Spain	1 = Spain, 0 if otherwise
AGE	AGE	Student age	Continuous
Grade	ST001D01T	International grade	Grade 7, grade 8, ..., grade 12 and ungraded
Ability	PV1MATH–PV10MATH	Student mathematics performance as a proxy for their academic performance. The variable PV1MATH is used in main analyses while others are for sensitivity checks.	Continuous
ESCSG	ESCS	Index of economic, social and cultural status as a proxy for family wealth. This variable is derived in the dataset based on highest parental occupation status, highest parental education, and home possessions.	1 = lowest status and wealth, ... , 10 = highest, and Missing
FAMSUPG	FAMSUP	Family support frequency based on students' rating of the frequency someone in their family engaged in activities with them outside of school.	High, Medium, Low, Missing
SKIP	SKIPPING	An indicator on whether the student had skipped classes in the two weeks before the survey.	Skipped, No skip, Missing
Gender	ST004D01T	Gender	1 = boys, and 0 = girls
WORKP	WORKPAY	Number of times students worked for pay outside of school in a typical week.	0 = none, 1 = once per week, ..., 10 = ten or more, Missing
WORKH	WORKHOME	Number of times students worked in the household or cared for a family member outside of school in a typical week.	0 = none, 1 = once per week, ..., 10 = ten or more, Missing
DFREN	WB158Q01HA	Number of days spent with friends after school.	0 = 0 days, ..., 6 = 6 days, Missing

Regression variable name	PISA variable name	Description	Variable values*
NFREN	WB156Q01HA	Number of close friends.	High, Medium, Low, Missing
RELATSTG	RELATST	Quality of relationship with teachers.	High, Medium, Low, Missing
BULLIEDG	BULLIED	Frequency of being bullied in the past 12 months.	High, Medium, Low, Missing
EXER	EXERPRAC	Number of times students exercised in a typical week.	0 = none, 1 = once per week, ..., 10 = ten or more, Missing
BMIG	STUBMI	Body mass index. This variable was only used in the model where body image perception was the outcome variable.	High, Medium, Low, Missing
STRATIOG	STRATIO	Student–teacher ratio as a proxy for school resources.	High, Medium, Low, Missing
SCHTYPEG	SCHLTYPE	School type as classified by PISA based on school type and sources of funding reported by school principals.	Private independent, Private government dependent, Public, Missing
NEGSCH	NEGSCLIM	Negative school climate. The extent of behavioural issues contributing to negative climate in the school the student was studying in at time of survey.	Severe, Medium, Low, Missing
CSRATIOG	RATCMP1	Computer–student ratio as a proxy for school resources.	High, Medium, Low, Missing

*Variables categorised as High, Medium, and Low (or Severe, Medium, and Low) were created by dividing students into three equally sized groups (unweighted) based on their values for the respective PISA variable. The High category includes students with higher values, and the Low category includes those with lower values. To avoid dropping a large amount of data from the regression models, an additional Missing category was created for many variables. Similarly, ESCS values were divided into 10 equally sized groups to form the ESCSG variable.

The impact of A Level exam scheduling on performance

Tim Gill (Research Division)

Introduction

A Levels are the most common post-16 qualification taken by students in English schools. Most A Level students in England take at least three subjects and most of these have three examined components, giving a total of nine exams, which are all taken at the end of the two-year courses. The A Level summer exam series usually runs for about 45 days, but most students will take all their exams in a considerably shorter number of days. As such, they are required to revise for and take a substantial number of exams in a short period of time. This is referred to in this article as their exam *load*. Because of this, students may have to make trade-offs as to where they concentrate their effort and may experience exam fatigue, particularly towards the end of the series. Thus, there is the potential for students' performance to be affected by their exam schedule.

There is limited previous research looking at the impact of exam scheduling on results. Goulas and Megalokonomou (2018) investigated this topic in the context of exams in Greece for 16-year-olds. They found an overall positive effect on performance of the number of previous exams, which may be due to a "practice" effect of taking more exams. In contrast, the number of days since the first exam was negatively related to performance, suggesting a possible fatigue effect (although this was limited to STEM[1] subjects). They also found a small negative effect of the number of days since the previous exam (for non-STEM subjects only).

Pope and Fillmore (2015) investigated how the number of days between Advanced Placement exams in the US affected performance. They found that students performed significantly better on the second exam when there was a longer gap. This was particularly the case for female and Asian students.

Thus, the limited prior research evidence suggests that there are several different ways in which exam scheduling might affect performance. In this research, we investigated three different mechanisms (as described in Goulas and Megalokonomou, 2018) through which this impact may be felt:

1. The number of days since the previous exam: if there is only a short gap between one exam and the next this leaves little time for revising for the second exam, which may impact performance.

1 Science, Technology, Engineering, and Maths.

2. The number of days since the first exam: towards the end of the exam series, students may experience exam fatigue, which could lead to lower performance.
3. The number of exams already taken: this may have a positive or negative effect on performance. There may be a practice effect whereby students who have taken more exams may perform better. On the other hand, students may experience exam fatigue if they have already taken many exams.

The main purpose of this research was to investigate whether exam scheduling impacts performance on A Levels through these mechanisms.

Data and methods

We used data from the Cambridge OCR awarding organisation (AO).[2] This is one of four AOs which offer A Levels in England. We restricted the data to the results of students who took at least three A Level subjects with OCR. This was because we were interested in exploring the impact of exam scheduling on those most likely to be affected (i.e., those taking many exams). As most students take three A Levels, including those taking fewer than three with OCR would include many students taking at least one A Level with another board (the results of which we did not have access to).[3]

We combined data from several different exam series (2016 to 2019 and 2023) into one dataset. Exams did not take place in 2020 or 2021 due to the COVID-19 pandemic and 2022 was not included as there were some adjustments made in that year (e.g., providing some advanced information on content included in exams) which made it unusual. In the earlier years (2016 to 2018), A Levels were in the process of being reformed. Therefore, the data in these years included some reformed subjects and some yet to be reformed. For the subjects which had not yet been reformed, exams could be taken in a modular fashion (i.e., at different time points throughout the two-year course). For consistency with later years, we only included students from these earlier years who took all their exams in the final exam series.[4]

The data included a variable indicating the date of each exam. For each student, the date of their first exam was recorded, and then for each subsequent exam the difference between the exam date and the first exam date was calculated. This was the "days since first exam" measure. The "days since previous exam" was calculated by simply subtracting the previous exam date from the current exam

2 This data was collected as part of the usual marking and processing of students' examination scripts and has been stored and used in line with Cambridge University Press & Assessment's Data Privacy notice (https://www.cambridge.org/legal/candidate-privacy-notice).

3 Ideally, we would have used data from all AOs, so that we could be certain that we had data on all exams taken by each student. However, the all-AO data available (e.g., National Pupil Database or GRADE data provided by the Department for Education) only includes grades or marks for the overall qualification. For our analysis we needed marks on individual components.

4 This was mainly students who took only exams in reformed subjects, but also included some students who took exams in pre-reform subjects, as long as they took all their exams in the final exam series.

date. Finally, a running total of the number of previous exams was calculated. An example of the data for one candidate is shown in Table 1. This shows one line per exam, ordered by exam date. This candidate's first exam was on 16 May. For this exam, the "days since previous exam" measure was set to Missing because this was their first exam. The "days since first exam" was 0. The second exam (23 May) was 7 days after the previous exam and 7 days after their first exam. They had taken one exam previous to this. Their final exam was on 21 June. This was one day after their previous exam, 36 days after their first exam, and prior to this they had taken 8 exams. The final column shows the marks achieved on each exam (as a percentage of maximum marks on the paper).

Table 1: Example student data

Subject	Component	Exam date	AM or PM	Days since previous exam	Days since first exam	No. of previous exams	Mark (as a % of max. marks)
Classical Civilisation	01	16/05	PM	.	0	0	84.0
Classical Civilisation	02	23/05	PM	7	7	1	82.7
Classical Civilisation	03	06/06	PM	14	21	2	88.0
Maths	01	06/06	PM	0	21	3	55.0
Biology	01	07/06	AM	1	22	4	52.0
Maths	02	13/06	PM	6	28	5	58.0
Biology	02	16/06	AM	3	31	6	50.0
Maths	03	20/06	PM	4	35	7	49.3
Biology	03	21/06	AM	1	36	8	69.6

This student also had two exams scheduled at the same time (PM on 6 June). The procedure when this happens is for the student to take one of these exams at the correct time and then the other exam in the other session on the same day (while being kept apart from other students).[5] As it was not possible to know which of these exams was sat first, we did not know the correct values for the "days since previous exam" or the "number of previous exams" variables. The way we dealt with this issue was different for the descriptive analysis and for the regression analysis. Which exam came first was not important for the descriptive analysis, so we chose one of the exams at random to be categorised as the first exam on the day (as shown in Table 1). However, for the regression analysis, we needed to know the order of the exams, and therefore we excluded all instances of exams scheduled at the same time.[6]

Descriptive analysis

The first part of the analysis was descriptive, focusing on the exam load on students. This includes the number of exams taken, the total time spent on exams, and the average amount of time between exams. Examples of common schedules and the most compressed schedules are also included. This analysis was undertaken at the candidate level (i.e., one row per candidate).

5 See page 24 of https://www.jcq.org.uk/wp-content/uploads/sites/2/2025/10/Print-version-JCQ-Instructions-for-conducting-examinations-2025_6_FINAL.pdf

6 The number of exams excluded was 5 433, approximately 5 per cent of the total.

Regression analysis

For the main analysis, that is, to investigate whether performance is affected by exam scheduling, regression models were fitted, predicting candidate performance (on individual exams) from the measures of exam scheduling and several contextual variables. This analysis was undertaken at the candidate exam level (i.e., one row per exam, per candidate).

The regressions fitted were multilevel models as these account for the hierarchical nature of the data. Specifically, three-level models were fitted, with results nested in centres, and centres nested in components. The clustering within centres accounted for the fact that results within a centre tended to be more similar than results between centres. The clustering within components accounted for the fact that centre results within a component tended to be more similar than results between components.[7]

The general form of the model was as follows:

$$y_{ijk} = \beta_0 + \beta_1 X_{ijk} + \beta_2 W_{jk} + \beta_3 Z_k + v_{0k} + \mu_{0jk} + e_{0ijk}$$

where y_{ijk} is the mark percentage for exam *i* in centre *j* and component *k*, X_{ijk} is the set of independent variables at exam level, W_{jk} is the set of independent variables at centre level, Z_k is the set of independent variables at component level, β_1 to β_3 are the sets of regression coefficients, v_{0k} is a component random effect, μ_{0jk} is a random effect of centre within component, and e_{0ijk} is the residual difference between actual and predicted mark.

Performance was measured by the mark achieved on the paper as a percentage of maximum mark. This meant that performance on components of different length (in terms of maximum mark) could be directly compared.

The independent variables of interest were the measures of exam scheduling. These were as follows:

- Days since previous exam:[8] This was converted from a continuous to a categorical variable. This was to allow for the effect of this measure in the regressions to be non-linear. We expect that the longer the time between exams, the better the performance is likely to be, because students will have more time to recover and to revise for the next exam. However, this benefit may tend to tail off as the number of days increases and this effect would not be captured by a continuous variable. The following categories were used: 0, 1, 2, 3, 4, 5, 6, 7, 8 or more, First exam. The "First exam" category was added because we wanted to include the first exam for each student in the analysis and without this category the "days since previous exam" measure would not have been applicable for these exams.

7 The Intraclass Correlation Coefficients (which measure the proportion of the total variance which is attributed to differences between clusters) were 0.20 for clustering within centres and 0.21 for clustering within components. These figures indicate that clustering was present and therefore a multilevel model was justified.

8 This measure takes no account of whether the exams were in the morning or the afternoon, so the true amount of time may be less or more than the full days stated. For example, if the number of days was equal to 1, this could actually be somewhat less than 24 hours if the previous exam was in the afternoon and the current exam in the morning.

- Days since first exam: A continuous measure, ranging from 0 to 45.
- Number of previous exams taken: As above, this variable was converted to a categorical variable, to allow for non-linear effects. Increasing the number of exams taken previously may be associated with improved performance to begin with (practice effect) but worse performance towards the end of the exams (fatigue effect). The categories were: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 or more.

It was not possible to include all exam scheduling variables in the same model, due to a high level of correlation (0.79) between the "days since first exam" variable and the "number of previous exams" variable. Therefore, two separate sets of models were fitted for different combinations of the exam scheduling variables, as follows:

1. "Days since first exam" and "days since previous exam".
2. "Number of previous exams" only.

The contextual variables in the models were as follows:

- **Exam series** – June 2016, 2017, 2018, 2019 and 2023.
- **Candidate gender** – Female or male, as recorded in the database.
- **Candidate concurrent attainment** – As measured by the Instant Summary of Achievement Without Grades (ISAWG) and recorded in the database. This measure (as described in Benton, 2017) is a summary, on a single scale, of the achievement of students taking OCR qualifications, no matter which qualifications were taken. A higher score indicates higher levels of attainment. Every student taking at least one OCR component will have an ISAWG score. This is a series-specific measure and therefore ISAWG scores cannot be directly compared across different series. As we were combining data from several different series, we needed to account for this in the models. To do this, we included an interaction term between the ISAWG and the series.
- **Centre type** – Centres were classified into one of: Comprehensive School, Independent School, Further Education (FE) College, Sixth Form College, Selective School, Other.
- **Subject group** – Each component was allocated to one of the following subject groupings: Creative, English, Humanities / Social Sciences, Information Technology, Maths, Professional, Sciences, Other. A list of which subjects were included in each subject group is shown in the Appendix.

For the regression analysis, continuous variables were centred around either their overall median value (days since first exam, days since previous exam, and number of previous exams) or their overall mean value (ISAWG).

Results

Descriptive statistics on the exam load

In this analysis, we examined the exam load for students taking A Levels. Table 2 presents descriptive statistics on the number of exams taken by each student in the OCR data, overall and broken down by exam year.

Note that there were far fewer students in the years 2016 to 2018 because we only included students taking all of their A Level exams at the end of the course.

Table 2: Statistics on the number of A Level exams taken

Year	Students	Mean	Median	Standard deviation	Minimum	Maximum
2016	127	11.9	12	3.0	6	18
2017	824	9.7	9	1.9	6	18
2018	1 424	9.5	9	1.8	6	20
2019	3 943	9.2	9	1.1	6	18
2023	4 429	9.4	9	1.3	6	18
All	10 747	9.4	9	1.4	6	20

The overall mean number of exams taken by students was 9.4, and the median was 9. Most students in the data took three A Levels and most A Levels consist of three examined components. The mean number of exams decreased between 2016 and 2019 but marginally increased in 2023. The maximum number of exams taken by a student was 18 in all years apart from 2018 where it was 20.[9] The minimum number was 6 in every year.[10]

Table 3 shows descriptive statistics on the amount of time (in hours) that students spent on their A Level exams, by exam series and overall.

Table 3: Statistics on the total amount of time spent on A Level exams

Year	Students	Mean	Median	Standard deviation	Minimum	Maximum
2016	127	19.11	19.50	4.16	9.25	28.50
2017	824	18.10	17.25	2.52	10.00	29.50
2018	1 424	17.78	17.25	2.67	11.00	32.00
2019	3 943	17.33	17.25	1.72	11.50	33.00
2023	4 429	17.68	17.50	2.04	12.00	31.75
All	10 747	17.61	17.25	2.12	9.25	33.00

On average, students spent over 17 hours on A Level exams. The shortest amount of time was just over 9 hours and the longest was 33 hours. The average amount of time spent decreased between 2016 and 2019 but slightly increased in 2023. This follows the same pattern as the average number of exams over time (see Table 2).

Table 4 and Figure 1 present statistics on the number of days between consecutive exams and Table 5 and Figure 2 show statistics on the number of days between the first and last exams for each student. These tables and figures give an indication of how compressed exam schedules tended to be and how long they tended to last.

9 These were students taking four or five A Levels with OCR. In each case, one of the subjects was Maths which allowed students to take multiple exams (more than the minimum necessary to obtain the qualification) and use their highest marks to contribute to the final grade.

10 These were students taking three A Levels, all of which had only two examined components.

Table 4: Statistics on the number of days between consecutive exams

Year	Students	Mean	Standard deviation	Minimum	Maximum
2016	127	3.32	0.89	1.60	6.20
2017	824	3.06	1.12	1.88	7.80
2018	1 424	2.35	0.74	1.25	6.17
2019	3 943	3.05	0.85	1.22	5.20
2023	4 429	3.41	0.83	1.70	5.86
All	10 747	3.11	0.92	1.22	7.80

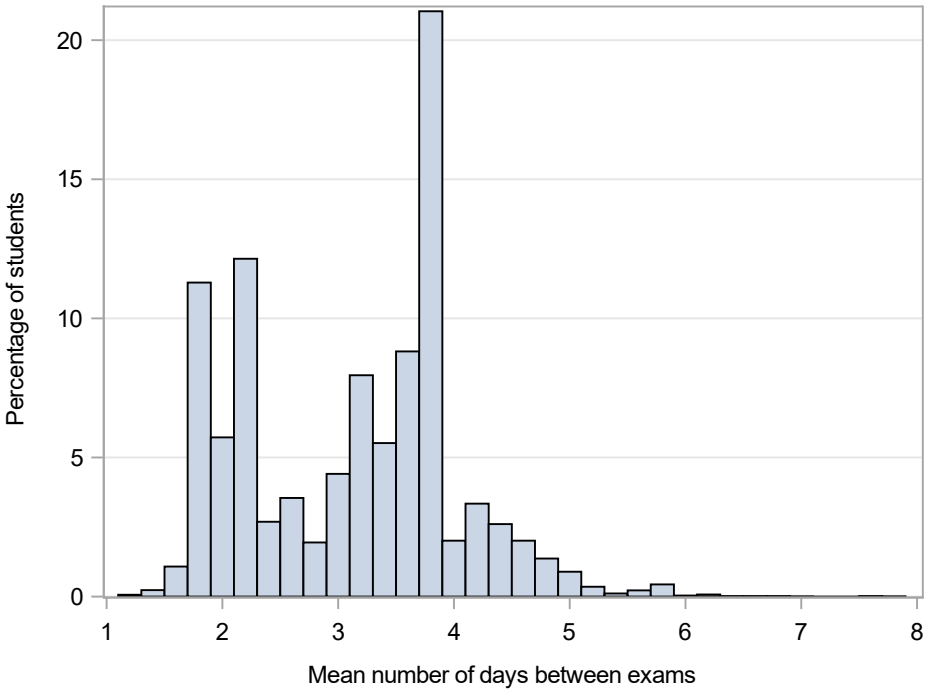

Figure 1: Distribution of mean number of days between consecutive exams

Table 5: Statistics on number of days between first and last exam

Year	Students	Mean	Standard deviation	Minimum	Maximum
2016	127	34.8	8.0	8.0	42.0
2017	824	26.6	11.3	13.0	45.0
2018	1 424	20.5	9.4	10.0	43.0
2019	3 943	24.8	6.6	11.0	42.0
2023	4 429	28.1	5.9	14.0	42.0
All	10 747	25.8	7.7	8.0	45.0

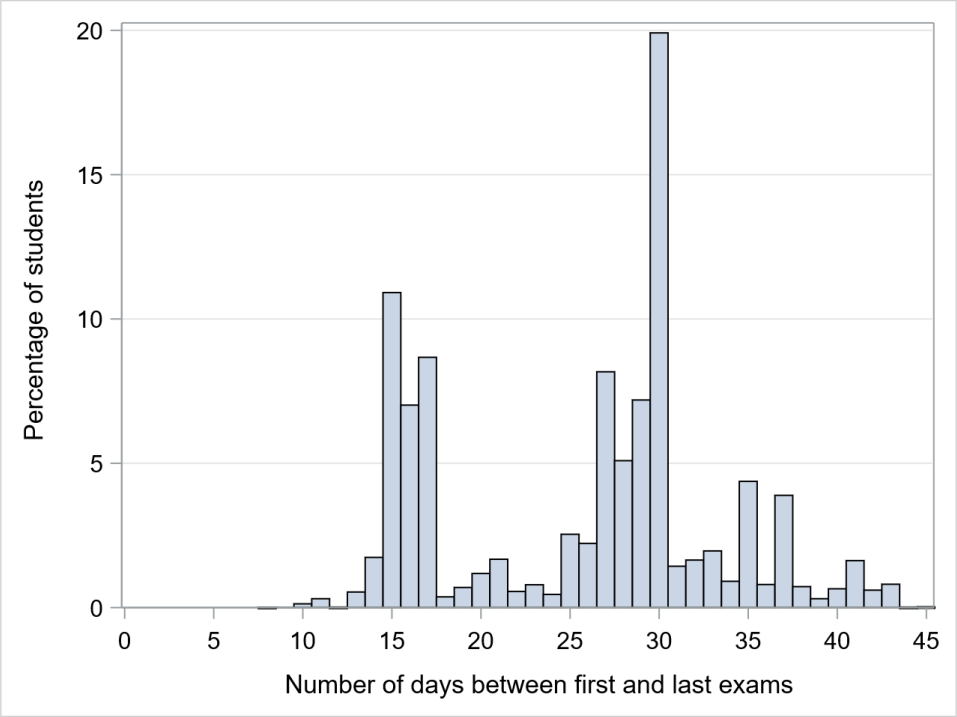

Figure 2: Distribution of number of days between first and last exams

On average, students had around three days between consecutive exams. There were some differences between years, with a mean of 2.35 in 2018 and a mean of 3.41 in 2023. The mean number of days between first and last exam was just below 26 days. Again, the lowest mean was for 2018 (20.5 days), with a considerably higher mean in 2023 (28.1 days). This suggests that students' exam schedules were much more compressed in 2018 than in 2023. This was at least partly due to changes implemented following the COVID-19 pandemic, whereby exams from the same qualification were scheduled further apart so that students had more time for revision (Department for Education, 2023).

The overall distributions (see Figures 1 and 2) for both these measures show two separate peaks, at around two days between consecutive exams (15 to 17 days between first and last exam) and at around 3.8 days between consecutive exams (30 days between first and last exam). This was mainly due to the half-term holiday in England, when schools were closed and no exams were taken. A significant minority of students had no exams before half-term and therefore only had around 15 to 17 days between the end of the half-term holiday and the end of the exam period (i.e., around two days between exams).

From the minimum values in Tables 4 and 5, it is clear that there were some students with very compressed schedules. We looked into these schedules in more detail. Table 6 shows the five most compressed schedules in terms of mean number of days between exams. This table shows the year in which the schedule took place, the number of students with that specific schedule, the mean number of days between consecutive exams ("Mean days"), the number of exams, the number of days between first and last exam ("Total number of days"), the total exam time and the subjects taken. For example, Table 6 shows that there were three students who took 10 exams (total time of over 19 hours) in 11 days in 2019 (an average gap of 1.22 days between exams). These students took Latin, Maths,

and Further Maths. There were also three students who took 13 exams in 15 days, with a total time of 25 hours.

Table 6: Five most compressed exam schedules (lowest mean days between exams)

Year	Students	Mean days	Number of exams	Total number of days	Total exam time	Subjects
2019	3	**1.22**	10	11	19 hrs 10 mins	Latin, Maths, Further Maths
2018	1	**1.25**	9	10	16 hrs 30 mins	Geography, Psychology, English Language and Literature
2019	3	**1.25**	13	15	25 hrs	Chemistry, Biology, Maths, Latin
2019	2	**1.31**	14	17	25 hrs 30 mins	Chemistry, Biology, Maths, Further Maths
2019	4	**1.33**	13	16	24 hrs	Chemistry, Biology, Maths, Further Maths
2018	1	**1.33**	13	16	25 hrs	Chemistry, Biology, Physics, Latin

We also looked at the most compressed schedule in terms of the least number of days to complete all exams. Using this measure, there were five students whose exam schedule was 10 days, in which they took either 9 or 8 exams.

Although the examples in Table 6 revealed some very compressed schedules, these were only for very few students. To put this in more context, we also looked at the most common exam schedule in each year. In all years apart from 2016, the most common schedule was very similar, requiring 9 exams (of around 17 hours total time) in between 15 and 17 days. The subject combinations taken by these students were either Biology, Chemistry, and Maths, or Biology, Chemistry, and Physics.

Regression results

Table 7 presents the results of the first set of regression models, which included the "days since first exam" and "days since previous exam" variables as predictors of performance.

In model 1 these two exam scheduling variables were included on their own and in model 2 contextual variables were added. Model 3 adds the statistically significant interactions between the exam scheduling variables and the contextual variables.

The results of model 1 show a statistically significant but very small negative effect (-0.10) of the number of days since the first exam. However, after including the contextual variables[11] (model 2), this effect disappears completely.

For the "days since previous exam" variable there were significant and positive effects in model 1 for all categories apart from 7 days and 8 or more days. In model 2, all categories were significant and positive. The reference category for this variable was 0 days, so these positive effects indicate that students performed better in exams taken at least a day after their previous exam than they did when the exam was on the same day as the previous exam.

11 Note that centre type did not have a statistically significant effect, so was removed from the models.

However, the size of the effect was very small in each case (between 1.65 and 3.51 per cent of maximum marks). To illustrate the size of the effect, Figure 3 shows the predicted mark percentages for the different levels of this variable, for a typical student.[12]

Table 7: Regression results (models including "days since first exam", and "days since previous exam" variables)

Fixed effect		Model 1 (n=95 129)	Model 2 (n=95 094)	Model 3 (n=95 094)
Intercept		61.63 (0.68)*	65.80 (2.17)*	66.11 (2.15)*
Days since first exam		-0.10 (0.01)*	0.00 (0.01)	0.01 (0.01)
Days since previous exam	0			
	1	2.29 (0.36)*	1.93 (0.22)*	1.38 (0.31)*
	2	1.25 (0.38)*	2.35 (0.23)*	2.30 (0.33)*
	3	2.78 (0.39)*	2.01 (0.24)*	1.67 (0.33)*
	4	3.20 (0.40)*	3.51 (0.24)*	2.68 (0.35)*
	5	1.87 (0.43)*	2.60 (0.26)*	2.71 (0.37)*
	6	3.74 (0.50)*	2.72 (0.30)*	2.88 (0.44)*
	7	-0.04 (0.70)	1.65 (0.42)*	1.61 (0.55)*
	8 or more	0.80 (0.41)	2.74 (0.25)*	2.13 (0.36)*
	First exam	1.97 (0.44)*	2.92 (0.27)*	2.36 (0.36)*
Gender	Female			
	Male		0.50 (0.08)*	-0.35 (0.37)
ISAWG			20.16 (0.43)*	17.70 (0.51)*
Subject group	Creative			
	English		7.50 (3.60)*	7.37 (3.56)*
	Humanities / Social Sciences		-0.30 (2.32)	-0.28 (2.29)
	Information Technology		-11.55 (4.15)*	-11.74 (4.09)*
	Maths		-14.23 (2.43)*	-14.06 (2.40)*
	Other		0.32 (2.60)	0.29 (2.56)
	Professional		2.40 (2.94)	2.07 (2.90)
	Sciences		-11.86 (2.46)*	-11.81 (2.43)*
Series	2016			
	2017		-1.71 (0.50)*	-1.80 (0.50)*
	2018		3.37 (0.51)*	3.34 (0.50)*
	2019		15.45 (0.51)*	15.41 (0.51)*
	2023		2.59 (0.51)*	2.47 (0.51)*
ISAWG*Series	2016			
	2017		2.47 (0.57)*	2.44 (0.46)*
	2018		2.06 (0.45)*	2.12 (0.45)*
	2019		4.87 (0.44)*	5.13 (0.44)*
	2023		6.82 (0.44)*	6.98 (0.44)*

12 For this exemplification and for others that follow, a "typical" student is one in the reference category for all applicable categorical variables (e.g., female, 2016 series, creative subject group) and with a value equal to the mean for all applicable continuous variables.

Fixed effect		Model 1 (n=95 129)	Model 2 (n=95 094)	Model 3 (n=95 094)
Days since previous exam*Gender	0			
	1			1.05 (0.39)*
	2			0.41 (0.41)
	3			0.75 (0.42)
	4			1.68 (0.44)*
	5			0.10 (0.47)
	6			-0.02 (0.55)
	7			-0.37 (0.68)
	8 or more			1.08 (0.45)*
	First exam			0.93 (0.43)*
Days since previous exam*ISAWG	0			
	1			3.81 (0.29)*
	2			2.95 (0.30)*
	3			2.78 (0.31)*
	4			2.83 (0.32)*
	5			2.74 (0.35)*
	6			1.79 (0.40)*
	7			-0.67 (0.52)
	8 or more			1.50 (0.33)*
	First exam			-1.91 (0.32)*

Note: Standard errors in parentheses. * $p < 0.05$.

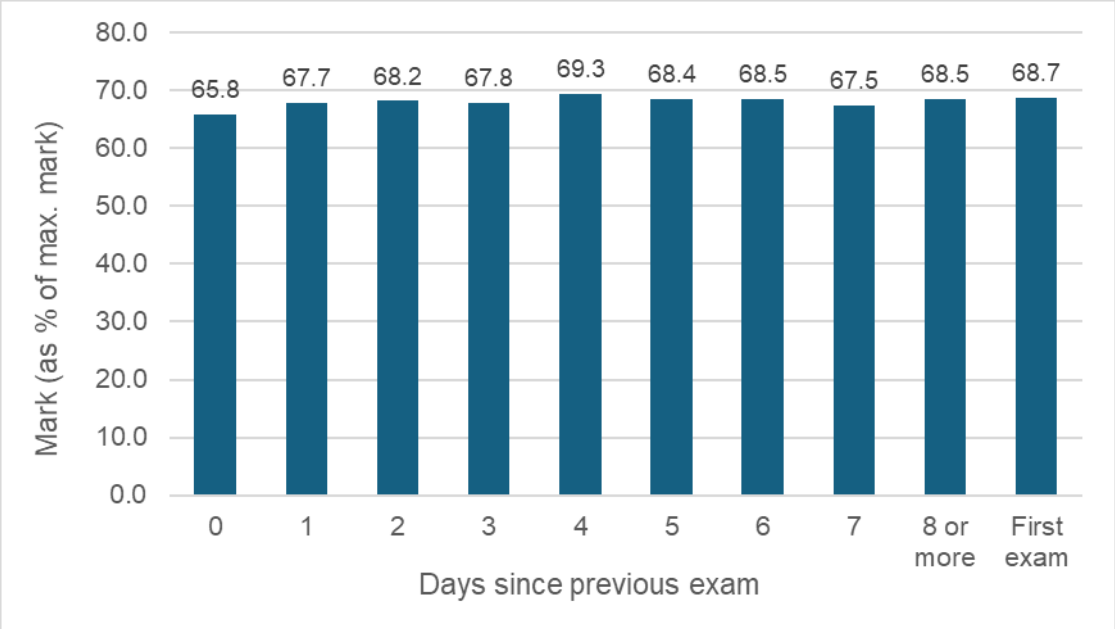

Figure 3: Predicted mark percentage for different levels of "days since previous exam" variable (model 2)

Figure 3 shows, for example, that when the days since the previous exam was 0 the predicted mark percentage was 65.8. With a break of 4 days the prediction was 69.3, an increase of 3.5 percentage points.

The results of model 3 (Table 7) show two significant interactions with the "days since previous exam" variable, for gender and for ISAWG score.

For the interaction with gender there were significant positive effects for 1 day, 4 days, 8 or more days, and First exam. This means that the positive effect of these numbers of days (compared with 0 days) was larger for males than females.

Figure 4 illustrates this effect for a "typical" student, showing the predicted mark percentage for males and females separately. The effects, however, were very small. For example, the largest difference compared to 0 days between exams for males was 4.4 percentage points for a gap of 4 days. For the same time between exams for females the difference was 2.7 percentage points.

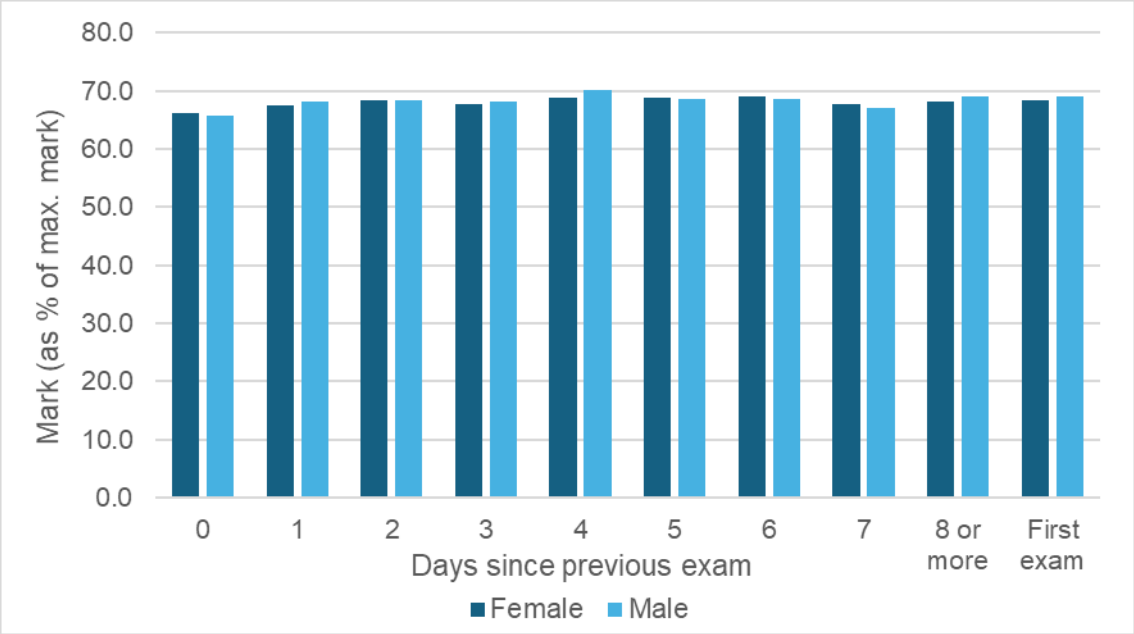

Figure 4: Predicted mark percentage by gender and "days since previous exam" variable (model 3)

The interaction effect with ISAWG score was positive (and significant) for all categories apart from 7 days (negative, but not significant) and First exam (negative and significant). This means that the positive effect (that is, better performance) of more days since the previous exam (compared with 0 days) was greater for those with higher ISAWG scores. However, the largest effect was for 1 day and for any more days than this the effects (although still positive) were lower. The interaction is illustrated in Figure 5, which shows the predicted mark percentage (for a typical student) for different numbers of days since the previous exam across a range of ISAWG scores. This figure shows that, for example, the largest difference in predicted mark percentage compared to 0 days for a student with an ISAWG score of 1.75 was for 4 days (5.4 percentage points). In contrast, for students with an ISAWG score of -0.25 the difference for 4 days was negative (-0.3 percentage points).

Table 8 presents the results of the second set of models, which included the "number of previous exams" variable. In model 4 the exam scheduling variable was

included on its own, model 5 added contextual variables,[13] and model 6 added the statistically significant interactions.

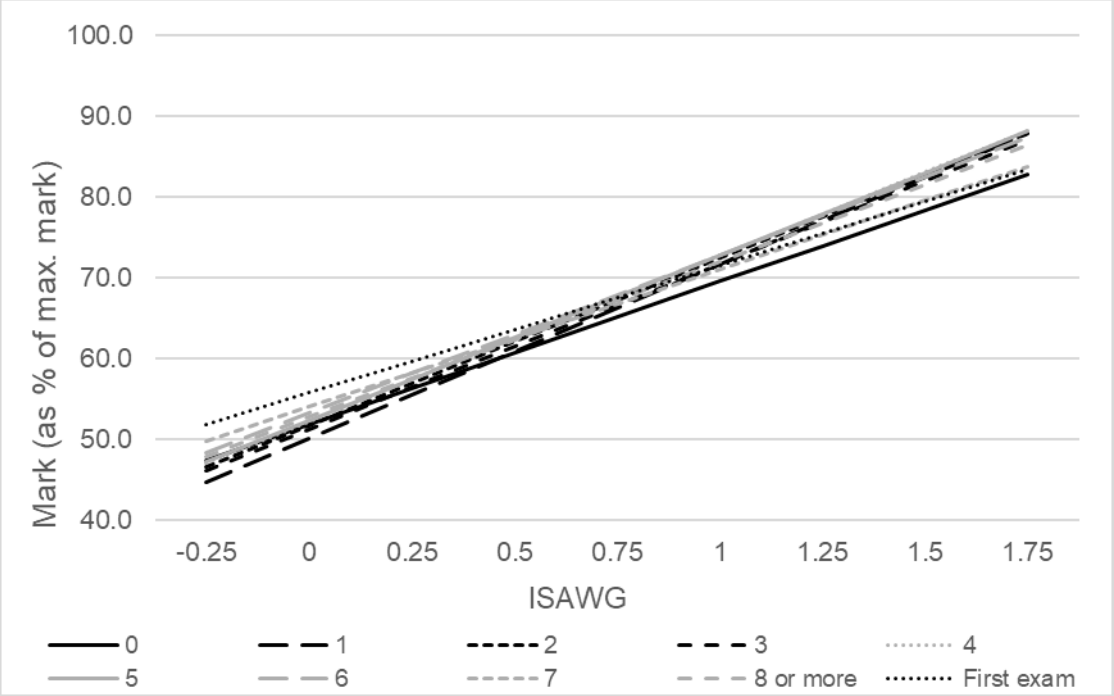

Figure 5: Predicted mark percentage for different levels of ISAWG and "days since previous exam" variable (model 3)

In models 4 and 5, there were small but significant negative effects for most categories of the "number of previous exams" variable. The results of model 5 indicate that students performed worse for each number of previous exams (apart from 2) than they did when the number was 0 (i.e., their first exam). Figure 6 shows the predicted mark percentages for the different categories of this variable, for a typical student. This shows that the differences in predicted performance were all very small, with a maximum difference of 2.5 marks (for 6 previous exams).

In model 6, the significant interactions between the number of previous exams and the contextual variables are included. There was only one significant interaction, with the ISAWG. This was positive for all numbers of previous exams and (mostly) increased with the number of previous exams taken. This means that the negative effect of more previous exams was lower for students of higher ability. Figure 7 illustrates this effect by comparing the mark percentage (for a typical student) for different numbers of previous exams and across a range of ISAWG scores. This shows that for high ability students there was a substantial positive effect of more previous exams, while for low ability students the effect was negative. The largest effect was for 9 or more previous exams. Students with an ISAWG score of 1.75 were predicted a mark percentage 7.1 percentage points higher for 9 or more previous exams compared with 0 previous exams. For students with an ISAWG score of -0.25 the predicted mark percentage for 9 or more previous exams was 11.2 percentage points lower than for 0 previous exams.

13 Note that centre type did not have a statistically significant effect, so was removed from the models.

Table 8: Regression results (models including "number of previous exams" variable)

Fixed effect		Model 4 (n=95 129)	Model 5 (n=95 094)	Model 6 (n=95 094)
Intercept		63.67 (0.68)*	68.82 (2.18)*	68.68 (2.13)*
Number of previous exams	0			
	1	-1.66 (0.31)*	-1.09 (0.19)*	-0.95 (0.19)*
	2	-0.19 (0.32)	0.25 (0.19)	0.40 (0.19)*
	3	-1.19 (0.34)*	-1.28 (0.21)*	-0.99 (0.21)*
	4	-1.88 (0.36)*	-2.04 (0.22)*	-1.83 (0.22)*
	5	-1.35 (0.37)*	-1.45 (0.23)*	-1.20 (0.23)*
	6	-2.31 (0.39)*	-2.52 (0.24)*	-2.27 (0.24)*
	7	-0.00 (0.42)	-1.34 (0.26)*	-1.30 (0.26)*
	8	0.70 (0.45)	-1.37 (0.28)*	-1.67 (0.28)*
	9 or more	5.55 (0.48)*	-0.03 (0.30)	-1.54 (0.31)*
Gender	Female			
	Male		0.51 (0.08)*	0.50 (0.07)*
ISAWG			20.14 (0.43)*	14.85 (0.46)*
Subject group	Creative			
	English		7.26 (3.57)	7.34 (3.54)*
	Humanities / Social Sciences		-0.48 (2.30)	-0.05 (2.28)
	Information Technology		-11.19 (4.11)	-11.01 (4.07)*
	Maths		-14.51 (2.41)	-13.92 (2.39)*
	Other		0.27 (2.57)	0.49 (2.55)
	Professional		2.52 (2.91)	2.87 (2.89)
	Sciences		-11.83 (2.44)	-11.32 (2.42)*
Series	2016			
	2017		-1.79 (0.50)*	-1.84 (0.50)*
	2018		3.26 (0.51)*	3.24 (0.50)*
	2019		15.40 (0.51)*	15.38 (0.50)*
	2023		2.53 (0.51)*	2.42 (0.50)*
ISAWG*Series	2016			
	2017		2.48 (0.47)*	3.20 (0.46)*
	2018		2.06 (0.45)*	2.92 (0.45)*
	2019		4.84 (0.44)*	6.06 (0.44)*
	2023		6.77 (0.44)*	7.71 (0.44)*
Number of previous exams*ISAWG	0			
	1			3.54 (0.22)*
	2			3.20 (0.22)*
	3			4.60 (0.22)*
	4			3.79 (0.22)*
	5			4.40 (0.22)*
	6			4.89 (0.22)*
	7			4.44 (0.22)*
	8			9.22 (0.24)*
	9 or more			9.15 (0.31)*

Note: Standard errors in parentheses. * $p < 0.05$.

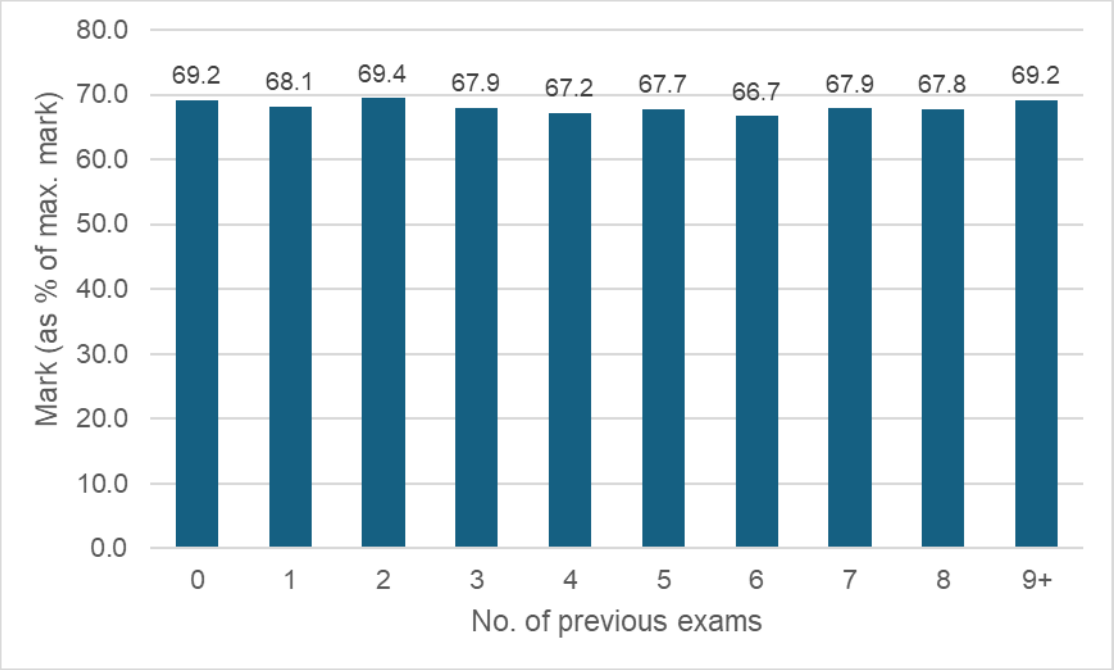

Figure 6: Predicted mark percentage for different levels of "number of previous exams" variable (model 5)

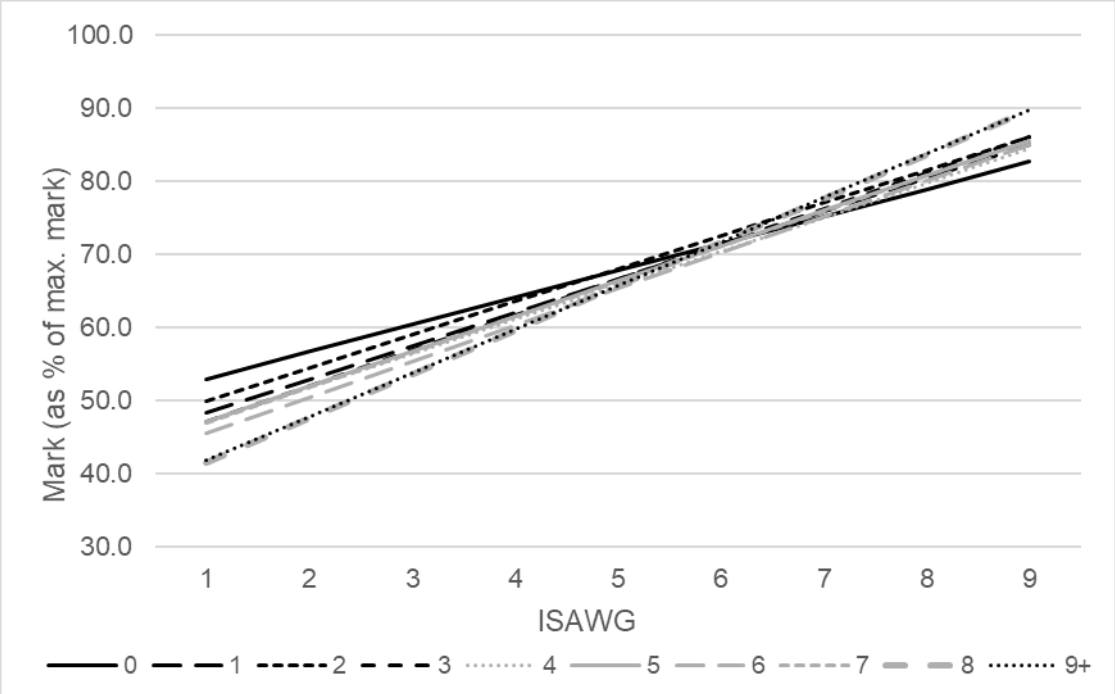

Figure 7: Predicted mark percentage for different levels of ISAWG and "number of previous exams" variable (model 6)

Discussion

This research investigated the current exam load on A Level students in England and the impact of exam scheduling on performance. The focus was on students taking OCR A Levels only.

In terms of the exam load, the median number of exams taken was 9 and the average total time was around 17.5 hours. The average number of days between exams was 3.1, but a substantial proportion of students had an average of around 2 days.

In terms of the impact of exam scheduling on performance, the results of the analyses showed no significant effect on performance of the number of days since the first exam (after accounting for gender, concurrent attainment, exam series, and subject). This means that there was no evidence of a fatigue effect, whereby students get tired or bored towards the end of the exam period. This finding contrasts with previous research (Goulas & Megalokonomou, 2018) which found a small negative effect for this variable, but for exams in STEM subjects only.

There was a significant and positive effect on performance of the number of days since the previous exam. However, this was essentially limited to any number of days being better than 0. There was no evidence that cumulatively more days was associated with higher performance. This suggests that students were less likely to do well on the second (or third) exam taken on the same day, compared with longer gaps between exams. Overall, this effect was very small (between 2 and 3.5 mark percentage points).

However, the effect was different for some groups of students. Firstly, it was slightly larger for male students than for females. Secondly, for students of higher ability there was a clear effect of improved performance for a gap of 1 day or more, but this was not the case for lower ability students. This finding is similar to that from previous research (Goulas & Megalokonomou, 2018) which found a positive effect of an increase in the number of days between exams for high ability students (but only on exams in non-STEM subjects). Their suggested reason for this was that high ability students may be better than low ability students at cramming for exams and therefore additional days between exams may be beneficial for these students.

For the final measure of exam scheduling (number of previous exams) there was a significant negative effect for most numbers of prior exams compared with it being the first exam. However, there was no evidence of larger effects for larger numbers of prior exams. Furthermore, the significant effects were all very small.

However, there was clear evidence of a differential effect of the number of previous exams on performance for students of different ability. For high ability students the more previous exams taken the better the performance. For low ability students the reverse was true. This effect was larger than the main effect, amounting to a 7-percentage point difference for students with an ISAWG score of 1.75 with 9 or more previous exams compared with their first exam. This effect may be due to different levels of motivation for low and high attainers. Low attainers may tend to be less motivated and therefore may be more likely to suffer exam

fatigue in later exams. In contrast, high attainers are usually more motivated and therefore more able to benefit from practice effects of taking more exams. Practice can potentially help with different aspects of exam taking technique such as time management, stress management and preparation strategies.

This finding was the opposite of previous research (Goulas & Megalokonomou, 2018), which found that, while the overall effect of the number of previous exams was positive, for higher ability students it was significant and negative. They argue that high ability students have stronger meta-cognitive skills and may therefore experience lower returns to practice. However, it is worth noting that the context of the previous study was different, involving students in 11th grade taking exams in an average of 13 different subjects in Greece.

In this research, we found no significant interaction effects between the measures of exam scheduling and subject group. This is in contrast to previous research (Goulas & Megalokonomou, 2018), which found differential effects depending on whether the exam was in a STEM or non-STEM subject. However, this could be because in this analysis, we used a more detailed breakdown of subjects, and the context was different in the previous study.

The overall findings from this research suggest that there is not a strong case for making substantial changes to the way that exams are currently scheduled. Nonetheless, a minor change which could be considered would be to try to minimise the number of occasions when candidates have two (or even three) exams on the same day, as candidates are slightly disadvantaged by this. This, however, may not be possible given the constraints involved in trying to schedule a large number of exams into a relatively short period of time.

A limitation of this research was the fact that we only had data from OCR. Inevitably, there would have been some students in our data taking three or more A Levels with OCR who also took one or more A Levels with a different board. For these students we did not know when they took their non-OCR exams, and therefore some of their exam scheduling data will have been inaccurate. We do not know how many students this applies to, but we can assume it is a small minority. For example, only 4.7 per cent of students took more than three A Levels in 2023.[14]

Another caveat is that by restricting to OCR exams only we do not know how representative this is of the population of all A Level students. The most common combinations of subjects in our data might be different from those for A Level students as a whole. However, according to Ofqual (https://analytics.ofqual.gov.uk/apps/Alevel/SubjectCombinations/), the most common combination in 2019 and 2023 was the same as we found (Biology, Chemistry, Maths),[15] which is a positive sign that our dataset shares characteristics with the wider cohort.

A final limitation relates to the specification of the multilevel model. We chose to nest results within centres and centres within components. An alternative

14 See https://www.gov.uk/government/publications/infographic-a-level-results-2023/infographics-for-a-level-results-2023

15 We could only compare subject combinations between our dataset and the wider A Level cohort for 2019 and 2023 due to the data available.

specification would have been to nest results within candidates, on the basis that results were more likely to be similar within candidates than between candidates. Not doing this clustering within candidates may have led to reductions in standard errors and hence increased the chance of variables being found to be significant. Further research could explore the possibility of using more complex multilevel models, which also account for the clustering within candidates.

Further research could also investigate in more detail the differential effects of exam scheduling on low and high achievers, to explore, in particular, why high achievers benefit from a practice effect while low achievers do not.

It would also be interesting to investigate the impact of the COVID-19 changes to scheduling, with exams from the same qualification being more spaced apart. In particular, did this have any effect on performance in the later exams because, for instance, the longer gaps meant candidates had more time to revise, or alternatively, candidates forgot what they revised for the previous exams?

Finally, further research could look at absence rates and whether they are higher towards the end of the exam period. This is another way in which exam scheduling could impact on results, but which was not captured by the analysis presented here.

References

Benton, T. (2017, November 9–11). *Pooling the totality of our data resources to maintain standards in the face of changing cohorts* [Conference presentation]. 18th annual AEA-Europe conference, Prague, Czech Republic.

Department for Education (2023, April 28). Exams in 2023 – everything you need to know. *The Education Hub*.

Goulas, S., & Megalokonomou, R. (2018). *Marathon, hurdling or sprint? The effects of exam scheduling on academic performance. IZA DP No. 11624.* Institute of Labor Economics.

Pope, D. G., & Fillmore, I. (2015). The impact of time between cognitive tasks on performance: Evidence from advanced placement exams. *Economics of Education Review, 48,* 30–40.

Appendix: List of subjects, by subject group

Subject group	Subjects
Creative	Design and Technology, Drama and Theatre, Film Studies, Music, Media Studies
English	English Language, English Language and Literature, English Literature
Humanities / Social Sciences	Ancient History, Business, Classical Civilisation, Critical Thinking, Economics, General Studies, Geography, Government and Politics, History, Psychology, Religious Studies, Sociology
Information Technology	Computing, Computer Science, Information and Communications Technology
Maths	Further Mathematics A, Further Mathematics B (MEI),[16] Mathematics A, Mathematics B (MEI)
Professional	Accounting, Health and Social Care, Law, Travel and Tourism
Sciences	Biology A, Biology B (Advancing Biology), Chemistry A, Chemistry B (Salters), Geology, Physics A, Physics B (Advancing Physics)
Other	Classical Greek, Dutch, French, German, Gujarati, Latin, Persian, Physical Education, Portuguese, Spanish, Turkish

16 Mathematics in Education and Industry.

Mode effects in digital versus paper-based exams and their relationship with item characteristics

Carmen H. J. Lim, **Carmen L. Vidal Rodeiro and Tim Gill** (Research Division)

Introduction

The digitalisation of high-stakes exams is becoming a reality. Some exams in Cambridge International AS and A Levels, as well as in Cambridge IGCSEs, will be available on screen for some schools from 2026.[1] There is also growing impetus for providing digital exams for some general qualifications in England.[2] This transition is driven by the benefits that on-screen assessments offer. For example, the availability of assistive technologies in digital exams makes assessments more accessible to students with additional needs. Furthermore, digital exams are more environmentally friendly than paper-based exams (Imran et al., 2021; Rogo Assessment, 2025) and allow certain skills, such as computer programming, to be assessed more authentically.

Exam boards may adopt several approaches to implement a digital exam, including transferring paper-based assessments to an on-screen format. Regardless of the taken approach, it is essential to maintain comparability across assessment modes. One issue of particular interest when digital and paper-based exams ask candidates the same questions is whether test items administered in the different modes (digital versus paper) would assess the same intended constructs and function similarly. Candidates may find an item harder or easier on screen than when responding to it on paper if the item assesses different constructs. In such instances, a mode effect is said to be present.

Existing research reports mixed findings about whether digital exams are harder or easier than paper-based exams. In fact, some studies show no mode effect, while others indicate better performance in either digital or paper-based formats.

1 For Cambridge International Education's timeline for digital exams, see https://www.cambridgeinternational.org/programmes-and-qualifications/developing-digital-assessment/digital-assessment-availability/.

2 https://www.gov.uk/government/collections/research-on-on-screen-assessment

For instance, using quasi-controlled empirical methods, Poggio et al. (2005) found very little difference in candidates' overall performance between the two modes in a large-scale mathematics test. Research from Li et al. (2017) and Steedle et al. (2020) showed that candidates tested on screen tended to perform better than those tested on paper in the ACT[3] test – a university admission test in the United States. This pattern was observed in all the test sections, including mathematics and science, but was especially visible in English (use of grammar), and in the reading and writing sections. On the other hand, Jerrim et al. (2018) analysed the 2015 Programme for International Student Assessment (PISA) test in reading, maths and science and concluded that students who took the paper-based version tended to perform better than those who took the equivalent digital version.

This heterogeneity in findings indicates that the issue is nuanced with the presence and direction of mode effects depending on a multitude of factors. Providers of digital exams may have used different testing platforms, each offering different functionalities. The design of on-screen items may have also varied across testing platforms, and the candidates included in these studies may have had different levels of experience with on-screen exams. These factors, among others, can contribute to differences in findings across studies. Therefore, continued research is essential to increase our understanding on this topic.

Mode effects can arise even when candidates are presented with the same questions on paper and on screen. For example, drawing on previous research, Jerrim (2016) noted that the cognitive processes required for reading can differ between paper and on-screen items, potentially influencing the approach and time taken to respond. Contributing factors may include candidates' digital competency which can either enhance or hinder their performance, and the presence of novel item types on screen which can increase candidates' engagement with digital exams. Furthermore, Hughes et al. (2011) found that candidates performed better on paper than on screen in a mathematics test, with the performance gap most visible for items requiring annotations on the provided diagrams or graphs.

To pave the way for digital high-stakes exams, Cambridge University Press & Assessment launched a Digital Mocks Service in 2023. This service provides candidates with valuable opportunities to familiarise themselves with digital exams by offering mock exams on screen. It is also an opportunity for schools to assess their readiness to offer digital high-stakes exams in the future. Since its inception, research has been conducted to continuously evaluate the delivery of the digital mock exams, including user experience and test validity (see, e.g., Kuvalja, 2023; Mistry & Kemmery, 2025; Vidal Rodeiro, Gill & Hughes, 2024).

One of the aims of the present work was to investigate the comparability between digital and paper-based exams. Specifically, we assessed the presence of Differential Item Functioning (DIF) by comparing the difficulty of exam items administered via the Digital Mocks Service and their equivalent paper versions. An item is said to exhibit DIF if two groups of candidates with similar ability do

3 https://www.act.org/content/act/en/products-and-services/the-act.html

not have the same likelihood of correctly answering the item. The presence of DIF does not necessarily indicate a mode effect, but it is an indication of unexpected functioning of the item that warrants further review (for example, asking relevant experts to review the item to consider if the difference in performance could be due to factors other than the assessment mode).

A second aim of this work was to increase our understanding of the potential factors driving mode effects. Specifically, we investigated whether items with a specific characteristic (e.g., requiring working with diagrams and graphs) are more likely to exhibit DIF, and hence, potentially have mode effects. To achieve this, we classified items based on their characteristics or features using an item classification framework developed as part of this research. This framework draws heavily on insights from Crisp and Ireland (2022), which proposed a structure for analysing features of digital items that may alter the assessed construct across testing modes.

The data included in our analyses is described in the next section, followed by the methods used for the DIF analysis and the item characteristics considered in the item classification framework. The findings of the overall DIF analysis and the frequency of DIF by item characteristics are presented next. The article concludes with a discussion and limitations section.

Data

We considered all the digital exams administered in the Digital Mocks Service from January 2023 to March 2025 and included those that met the following criteria:

1. The digital exam had an equivalent paper-based version that was used in a live session. This restriction is required as the fundamental principle of the analyses is to compare candidates' performance on items delivered both digitally and on paper. Digital exams based on specimen or practice papers which were not taken as part of a live exam session by any candidate were, therefore, excluded.
2. Both the digital and paper versions of the exam had at least 100 entries. This restriction ensures that our findings are not based on small samples of candidates.

In addition, we excluded one exam where a participating school was known to have experienced connectivity issues during the digital exam. The entries from that school accounted for about half of the total entries in that exam. Removing the entries from the affected school would have left us with fewer than 100 entries. While this issue reflects the inherent practical challenges of administering a digital exam, the data had to be excluded as any potential differences in candidate performance across modes might not be an indication of the functioning of the items.

In a small number of instances, a particular live exam paper was used as the basis for a digital mock exam in more than one digital mock session (e.g., the live exam for GCSE Computer Science Paper 1 from summer 2022 was available as a digital mock exam in both 2024 and in 2025). For such instances, the same digital exam used in two different digital mock sessions was treated as two separate exams

in our analysis. This approach avoids assuming that the digital interface was identical across sessions.

Throughout this report, the term "item" is used to refer to a question, or a sub-question when a question includes multiple parts. Furthermore, where the same live exam paper was used as a digital mock exam in two sessions, the parallel items in the two digital sessions were treated as separate items in our analysis.

Table 1 below presents the number of exams, number of items and entries analysed, by qualification.

Table 1: Number of exams, items and entries analysed, by qualification

Qualification	Number of exams	Number of items	Total number of entries	
			Digital mock	Paper live
GCSE	9	225	2 858	510 336
IGCSE	16	465	7 961	139 950
International AS or A Level	6	105	4 642	150 409
Total	**31**	**795**	**15 461**	**800 695**

Item-level data from the mock exams and from the equivalent paper versions was obtained from the Digital Mocks Service and Cambridge University Press & Assessment data systems, respectively.[4]

A wide range of subjects was included in the data. These fall broadly into 10 areas: Computer Science, English Language, Global Perspectives, Accounting, Economics, Biology, Physics, Chemistry, Psychology, and History. For English Language, the exams consist of reading and writing components. For Accounting, Economics, Biology, Physics, and Chemistry, the exam papers consist only of multiple-choice items.

Methods

Differential Item Functioning

The fundamental idea behind Differential Item Functioning (DIF) is that the difficulty of an item should be the same for any two groups of candidates, once their ability is taken into account. If an item is estimated to be harder for one group of candidates than for another group with similar ability, then the item is said to exhibit DIF.

There are many methods to assess DIF and, in this article, we used a well-established and effective approach: Item Response Theory (IRT) (see, for example, Thissen et al., 1993).

4 This data was collected as part of the usual marking and processing of students' examination scripts or as part of the marking and processing of digital mock exams, and has been stored and used in line with Cambridge University Press & Assessment's Data Privacy notice (https://www.cambridge.org/legal/candidate-privacy-notice).

Specifically, and due to the exams considered here including a mixture of dichotomous and polytomous items, we estimated the difficulty of each item in each mode using Partial Credit Models,[5] fitted using Winsteps (Linacre, 2023). The difficulty estimates produced using Winsteps have explicitly ensured the difficulty of the same item across modes can be meaningfully compared (that is, the difficulty estimates were put on the same scale). In these models, the candidate ability was approximated by the total marks scored in the exam.[6] Item difficulty is expressed in logits, where a value of 0 represents the average item difficulty.

Once we had estimated the difficulty of each item in both digital and paper modes, we checked if the differences in difficulty between modes were statistically significant at the 5 per cent level. For items where the differences were statistically significant, we then calculated the size of the difference between modes in absolute terms to classify them into three categories (henceforth, DIF size):

1. **Negligible,** where the difference in difficulty between modes is less than 0.43 logits
2. **Slight to moderate,** where the difference in difficulty is greater than or equal to 0.43 logits and less than 0.64 logits
3. **Moderate to large,** where the difference in difficulty is 0.64 logits or above

These threshold values correspond to those used by the Educational Testing Service (Zwick et al., 1999).

An item is considered to exhibit DIF if there is a statistically significant difference in the difficulty estimates between the digital and paper modes, and the DIF size is in the "slight to moderate" or "moderate to large" category. Defining items with DIF this way ensures that differences in difficulty have practical significance and are not likely to have occurred by chance.

Item classification framework

The next step in our research involved investigating whether items with specific features and characteristics were more likely to show DIF.

Instead of just focusing on items exhibiting DIF, we reviewed all items. Doing so helps to avoid confirmation bias[7] and provides us with a better understanding

5 For a detailed description of the Partial Credit Model (PCM), see De Ayala (2009). In one instance where the exam paper included optionality (which led to a small number of responses for several items), we used a Rating Scale Model (RSM) instead of a PCM. In a PCM, when the sample / cohort sizes are small, the estimation of the difficulty of the items with only a few, or maybe no observations in some mark points (at least 10 observations per mark point are recommended), will not be accurate. As the RSM is not as sensitive to this issue as the PCM, the RSM, also fitted using Winsteps, was better suited to analyse data for this exam paper. Details about comparing and choosing between a PCM and a RSM are given in Linacre (2000).

6 Considering an alternative ability measure (e.g., any other measure that evaluates the construct of interest) could have accounted for the possibility that the overall exam may exhibit a mode effect. However, in this research, an external measure of candidates' ability was not available and ability was therefore approximated by the total marks in the exam.

7 Confirmation bias, in this case, means the tendency to search for item characteristics that support the researchers' prior knowledge regarding the relationship between mode effects and item characteristics.

of whether items with certain characteristics are more likely to exhibit DIF (and hence, mode effects). For instance, do items requiring annotations have a higher frequency of DIF compared to items with other characteristics?

To do this systematically, we created an item classification framework that included a list of characteristics to screen each item for. This framework was primarily adapted from work by Crisp and Ireland (2022) but it also included a few additions such as the types of reading skills required when interacting with texts (e.g., retrieve versus reflect), which were considered by Harrison et al. (2023). Two of the researchers independently considered each item and noted their characteristics as covered in this framework. Any discrepancies were subsequently discussed to reach consensus.

The framework has five broad categories, each with multiple subcategories, totalling 33 characteristics. A brief overview of each characteristic category is discussed below. Table 2 includes the complete set of 33 characteristics, alongside their definition and examples. Note that some of the examples have been created to illustrate the subcategory and do not necessarily represent items in the exams included in the research.

The five broad characteristic categories are:

1. **Presentation and navigation:** This category captures the differences between having the same item on paper and on screen, in terms of presentation and functionality. The presentation of an item, such as the layout, response format, or the size and nature of the answer space, might change when presented digitally in ways that may affect how candidates understand and answer the same question in different modes. Additionally, a candidate may need to navigate differently on screen to read the resource and question and may have access to writing support features which could result in a different test-taking experience than that for paper-based candidates.

2. **Stimulus:** This category identifies the inclusion of text-heavy resources alongside the item and of other non-text materials such as graphs and complex tables. There is a lot of potential for differences in presentation of item stimuli across modes (e.g., text materials or tables might look different due to screen shape, screen size, or their position on the page). However, even if they were visually identical, previous research has argued that the required cognitive effort to process the information embedded in these materials, including reading-intensive passages, can be different across modes (for example, Fishbein, 2018).

3. **Response format:** This category includes a list of potential response formats on screen. At least one response format subcategory was assigned to each item based on their digital version. The inclusion of this category allowed us to investigate whether items with a particular response format are more likely to exhibit mode effects. It also helped capture response formats specific to digital exams, such as “drag and drop”, to understand whether they are likely to simplify or complicate the answering process.

4. **Response strategy:** This category captures how a candidate might have approached the item to arrive at an answer. This includes whether an item requires calculation, requires working with graphs, or relates to the answer from a previous item. Even if the required response strategy is technically the same across modes, how candidates carried it out can differ in practice since not all actions that are feasible on paper are feasible on screen, and vice versa. For instance, it would be harder for candidates to annotate a graph in a digital item with the same level of ease and to the same extent as annotating it on paper. Since our analyses included several exams with long reading passages, we also considered three aspects of reading literacy that indicate "the underlying cognitive processes that test takers are expected to utilise in answering items", as provided in Harrison et al. (2023, p. 14).

5. **Tariff:** This category intends to group items based on their maximum attainable marks into three categories: low, medium and high tariff. It is unlikely that item tariff itself would drive mode effects. However, high-tariff items tend to have relatively more complex characteristics than low-tariff items, which could collectively amplify any potential mode effect. Therefore, the inclusion of this category is, to some extent, an attempt to capture any interaction effects between multiple categories.

Table 2: Item classification framework

	Code	Characteristic	Description	Examples
Presentation and navigation	A1_1	Has at least one change to presentation, format or layout with potential impact on the ease of reading	An item's presentation, format or layout is different across modes in ways that could impact candidates' understanding of or time taken to read the item.	Text in grey font rather than the higher contrast black font; inconsistent line spacing.
	A1_2	Has at least one change to presentation, format or layout relating to the response method	An item's response method is different across modes in ways that could demand additional computer skills or require familiarity with the response method.	Drop-down lists in place of ticking a box; drafting computer code using a text editor instead of hand-written code.
	A1_3	Has at least one change to the size or nature of the answer space	The structure or size of the answer space is different across modes, which may lead candidates to have different expectations about response length and style.	Answer boxes are the same size on screen regardless of mark tariff, but on paper, they are smaller for one-mark items than for two-mark items.
	A1_4	Requires navigation to see the provided resources in full	Scrolling or switching between pages is needed to read the provided resources in full, which could, for example, cause content to be partially overlooked.	The two resources for an item cannot be displayed on screen simultaneously.
	A1_5	Requires navigation to see the whole question and the answer spaces	Scrolling or switching between pages is needed to read the question in full, which could, for example, unintentionally assess other constructs such as attention to detail.	A need to scroll back to the information at the start of a question when answering later sub-questions.
	A1_6	Contains writing support not available on the paper version	On-screen writing support could either help or hinder writing performance compared to when answering the same item on paper.	Automatic word count helps monitor writing progress relative to candidates' target answer length.
	A1_7	At least one change in the instruction or command words	The instruction given within an item might change in the digital version to match other changes in the test or item. This change risks learners interpreting the instruction and item differently across modes.	Changes to instruction to reflect a change in response format (e.g., "select an image ..." instead of "draw this ...").
Stimulus	A2_1	Heavy reading demand	The item has a text-heavy passage that is at least an A4 page long on paper.	-
	A2_2	Has at least one graph or diagram	The item has at least one graph or diagram of any complexity. Pictures (non-technical, for pure illustration purposes) are not considered here.	-
	A2_3	Has at least one complex table	The item has a table with at least five columns or rows, or a smaller table heavily populated with text.	-

	Code	Characteristic	Description	Examples
Response format	A3_1	Multiple choice	Requires selecting one correct answer from multiple options.	Choose one answer between options A, B, C and D.
	A3_2	Multiple response	Requires selecting more than one correct answer from multiple options.	Choose two answers between options A, B, C and D.
	A3_3	Numeric or maths entry	Requires providing a numeric value or mathematical symbols as the answer.	Enter a number between 1 and 10.
	A3_4	Inline choice	Requires selecting answers from a drop-down list.	Select an answer from the drop-down list.
	A3_5	True / False choice	Requires selecting between "True" and "False" as the answer.	Is this statement true or false?
	A3_6	Matching / pairing	Requires matching values or texts in rows or columns to form an answer.	Match each value to one of the provided texts.
	A3_7	Drag and drop	Requires dragging elements to a predefined area, such as a box, or on an image or within texts, to form an answer.	Click and drag the images to the correct gaps.
	A3_8	Text entry	Requires providing texts as answers, ranging from a single word to a short-constructed response with several sentences.	Give one definition of a syntax error in a program.
	A3_9	Hotspot	Requires indicating an answer by clicking on a point or an area inside an image.	Click on the image to indicate where you think the septum of a heart is.
	A3_10	Code compile	Requires typing a code snippet as an answer.	Complete the function *position()*.
	A3_11	Extended response	A constructed-response item worth at least six marks.	Write a story about the life of a cat.
	A3_12	Slider	Requires moving an indicator on a bar to select an answer.	Drag the bar to indicate your answer.
	A3_13	Graph and diagram	Requires providing a new graph or diagram as an answer.	Provide a flowchart.

	Code	Characteristic	Description	Examples
Response strategy	A4_1	Requires calculation	Requires calculation of any complexity, which could lead to candidates approaching the item differently depending on the mode.	What is 4 divided by 3?
	A4_2	Requires working with graphs or diagrams	Requires using information in a diagram or graph to arrive at an answer, which could be harder on screen if it involves sketching and annotating.	The graph shows the motion of a vehicle. What is the distance between X and Y?
	A4_3	Requires manipulating visuals or objects	Items where moving the provided visuals or objects from their original position or angle could help or hinder answering the item.	Order the options to arrive at the correct sequence.
	A4_4	Relates to responses from a previous item	Response depends on the given answer on the previous item, which may require additional navigation or recall from candidates.	Explain the reason for your choice in part (b).
	A4_5	Requires candidates to "access and retrieve" information from texts	A reading item requiring using content primarily within the given texts, focusing on the specific pieces of information, to arrive at an answer.	What are the benefits of exercising according to the text?
	A4_6	Requires candidates to "integrate and interpret" information from texts	A reading item requiring using content primarily within the given texts to form a broad understanding and to develop an interpretation about the topic.	The article provides the benefits of jogging and swimming. Give two reasons why swimming may not be appealing to some people.
	A4_7	Requires candidates to "reflect and evaluate" information from texts	A reading item requiring reflecting and evaluating the given content by connecting them with knowledge outside the provided texts.	Using your own words, explain what the text means by "indifferent".
Tariff	A5_1	Low-tariff items	One-mark item.	-
	A5_2	Medium-tariff items	Items worth two to five marks.	-
	A5_3	High-tariff items	Items worth at least six marks.	-

Results

Overall frequency of DIF

Table 3 presents the number and percentage of items with Differential Item Functioning (DIF) by qualification and for all qualifications combined. As can be seen in the table, 18 per cent of items (i.e., about one in five) were found to exhibit DIF (that is, the difference in difficulty between modes was statistically significant and categorised as "slight to moderate" or "moderate to large"). The frequency of DIF was very similar across qualifications, ranging narrowly from 17 per cent to 19 per cent.

The last two columns of the table show the number of DIF items estimated to be harder in each mode. Overall, 69 items were estimated to be harder on screen and 76 on paper which is roughly an equal split between modes. This shows that items found to potentially have a mode effect were not uniformly harder on screen or on paper. There were, however, some variations across qualifications. For International AS or A Level in particular, the split was roughly 30 per cent of DIF items harder on screen and 70 per cent harder on paper. That said, since this result is based on only 20 DIF items, these percentages could change noticeably if the difficulty of a DIF item shifts in the opposite direction. The results across qualifications should not be compared, as each included a different set of subjects and, therefore, items with different characteristics.

Table 3: Summary of overall Differential Item Functioning (DIF) results

Qualification	Number of items	Number of DIF items	Percentage of DIF items	Number of items harder for...	
				Digital	Paper
GCSE	225	38	16.9	14	24
IGCSE	465	87	18.7	49	38
International AS or A Level	105	20	19.0	6	14
All	**795**	**145**	**18.2**	**69**	**76**

Figure 1 depicts the estimated difficulty for each item on paper and on screen with each point on the graph representing an item. The figure also shows whether DIF was detected for the item and the size of the DIF if present. Any point that falls on the equality line (black solid line) was estimated to have the same difficulty across modes.

The figure shows an additional insight on top of the findings from Table 3: the presence of DIF is spread out across the item difficulty range. This implies that items with, say, higher demand, which are usually included in an exam with higher-ability candidates in mind, are not necessarily more susceptible to mode effects. In other words, even if a mode effect exists for these items, it did not disproportionately disadvantage candidates based on their ability level.

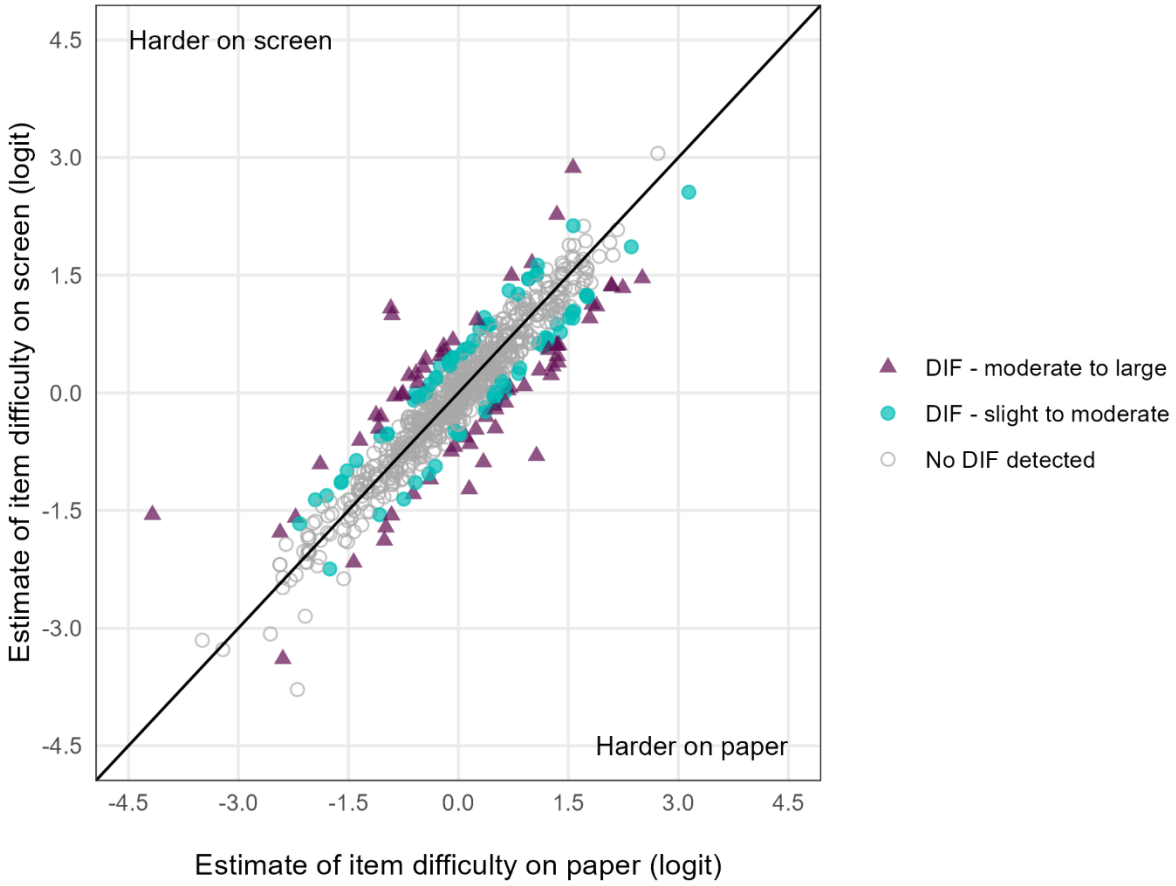

Figure 1: Estimated difficulty of each item on paper and on screen, and indications of the presence and size of Differential Item Functioning (DIF)

DIF by item characteristics

The number of items as well as the number and percentage of items with DIF in each item characteristic subcategory is presented in Table 4.

The first observation to note from the table is that we did not identify any items for four characteristics. These characteristics, which relate to the response format category, were: "True / False", "matching / pairing", "hotspot" and "slider". Similarly, several characteristics had very small numbers of items. Examples included items with a drop-down list as the response format, items that require visual manipulation, and items where responses are related to the previous item. Statistically speaking, findings based on a large number of items tend to be more reliable than those produced using fewer items. Thus, findings for characteristics with a low number of items should be interpreted with caution.

Table 4: Summary of results for Differential Item Functioning (DIF) by item characteristics. "N" denotes the "number of".

Code	Short description	Total N items	N DIF items	Percentage items with DIF	N items harder for...	
					Digital	Paper
A1_1	Change impacting reading	186	39	21.0	18	21
A1_2	Change in response method	31	2	6.5	1	1
A1_3	Change in answer space	46	6	13.0	4	2
A1_4	Needs navigation to read resource	169	27	16.0	9	18
A1_5	Needs navigation to read question	161	34	21.1	15	19
A1_6	Contained writing support	88	15	17.0	9	6
A1_7	Change in instruction	102	6	5.9	2	4
A2_1	Heavy reading	136	21	15.4	7	14
A2_2	Includes graph or diagram	81	16	19.8	8	8
A2_3	Includes complex table	118	22	18.6	10	12
A3_1	Multiple choice	368	68	18.5	40	28
A3_2	Multiple response	3	0	0.0	0	0
A3_3	Numeric or maths entry	44	13	29.5	3	10
A3_4	Drop down	9	1	11.1	1	0
A3_5	True / False	0	-	-	-	-
A3_6	Matching / pairing	0	-	-	-	-
A3_7	Drag and drop	11	-	-	-	-
A3_8	Text entry	267	54	20.2	16	38
A3_9	Hotspot	0	-	-	-	-
A3_10	Code compile	23	4	17.4	4	0
A3_11	Extended response	80	8	10.0	5	3
A3_12	Slider	0	-	-	-	-
A3_13	Graph or diagram	7	1	14.3	0	1
A4_1	Requires calculation	67	13	19.4	7	6
A4_2	Requires working with graphs / diagrams	20	2	10.0	0	2
A4_3	Requires visual manipulations	6	1	16.7	0	1
A4_4	Relates to previous item	8	3	37.5	1	2
A4_5	Access and retrieve text information	48	9	18.8	3	6
A4_6	Integrate and interpret text information	64	7	10.9	5	2
A4_7	Reflect and evaluate text information	54	12	22.2	3	9
A5_1	Low tariff	464	95	20.5	47	48
A5_2	Medium tariff	234	42	17.9	17	25
A5_3	High tariff	97	8	8.2	5	3

Figure 2 visualises the relationship between the number of items with each characteristic and the percentage of items with DIF. Each point on the graph represents a specific characteristic and the characteristic code is labelled alongside it. The short definitions of these codes can be found in Table 4 or see Table 2 for expanded descriptions.

In Figure 2, we also overlaid a horizontal dotted line to indicate the frequency of DIF averaged across all the items (18 per cent), which was presented earlier in Table 3. We would expect characteristics with numbers of items closer to the total number of analysed items to have a percentage of DIF items close to the overall DIF frequency. This can be seen in results for several characteristics such as one-mark items (A5_1) and items that are multiple choice (A3_1) which were the more common characteristics in our data.

We also overlaid a vertical dotted line in Figure 2 to indicate a threshold of 30 items. There are currently no established guidelines on the minimum number of items required for this type of analysis. Therefore, any selected threshold is inherently arbitrary. Our choice of 30 items is intended to offer readers a reference point for exercising caution when interpreting results. This threshold was selected based on the commonly accepted statistical practice that having at least 30 observations is considered sufficient for a reliable analysis.

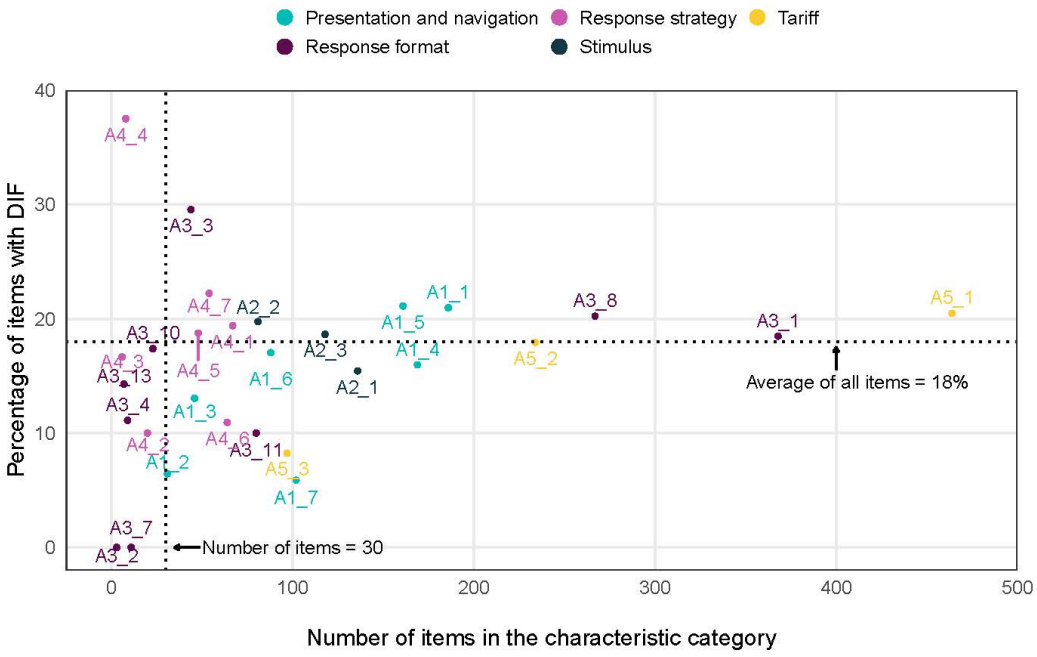

Figure 2: Number of items with each characteristic and the percentage estimated to exhibit Differential Item Functioning (DIF). Due to limited space, only characteristic codes are shown here; refer to Table 2 for full descriptions.

The key observation from Figure 2 is that the percentage of items with DIF for each characteristic was close to 18 per cent (the overall DIF frequency) or below for most characteristics. There were five characteristics with slightly more than 20 per cent of items with DIF but the percentage was not concerningly high. In ascending order, these were: items requiring text entry (A3_8); one-mark items (A5_1); items where there were some changes in presentation on screen that may

affect the ease of reading (A1_1); items where navigation is required to read the items in their entirety (A1_5); and items that require candidates to reflect and evaluate text information from passages (A4_7).

However, there were two characteristics where the detected frequency of DIF was higher. These were: items requiring numeric or maths entry as answers (A3_3); and items where the response relates to the previous item (A4_4). The frequency of DIF was particularly high for the latter, but we cannot be confident that this finding is meaningful as it was based only on eight items.

The results for items requiring numeric or maths entry were, on the other hand, based on many more items (44 items). Around three in 10 items were detected to have DIF, which was more than the DIF frequency detected for other characteristics with a similar total number of items.

Figure 3 focuses on the DIF items and illustrates the percentage of DIF items that were estimated to be harder in each mode, for each item characteristic. This result should be interpreted alongside the number of DIF items detected in each characteristic, as many of them only have a small number of DIF items. The characteristics were ordered from top to bottom in the graph by the number of DIF items detected, with those having a higher number at the top.

Following on from the earlier observation relating to items requiring "numeric or maths entry" (A3_3), it can be seen from Figure 3 that among the 13 DIF items detected, 10 were found to be harder on paper and three on screen (77 per cent versus 23 per cent, respectively).

There were several characteristics where all the DIF items were found to be harder either on paper or on screen. However, these results were based on four items at most; hence, they should be interpreted with caution.

Lastly, if we focus on characteristics with at least 30 DIF items, we can see that the direction of the DIF appears to be roughly evenly split between the two modes (or at most a 60:40 split). The only exception was items requiring text entry as answers (A3_8; not extended responses), where 70 per cent of the DIF items were found to be harder on paper.

Overall, our findings showed that the percentage of items with DIF in each of the different item characteristic categories had a wide variation. However, among the characteristics for which we have sufficient data to draw a confident conclusion, almost none of the findings stood out as concerning.

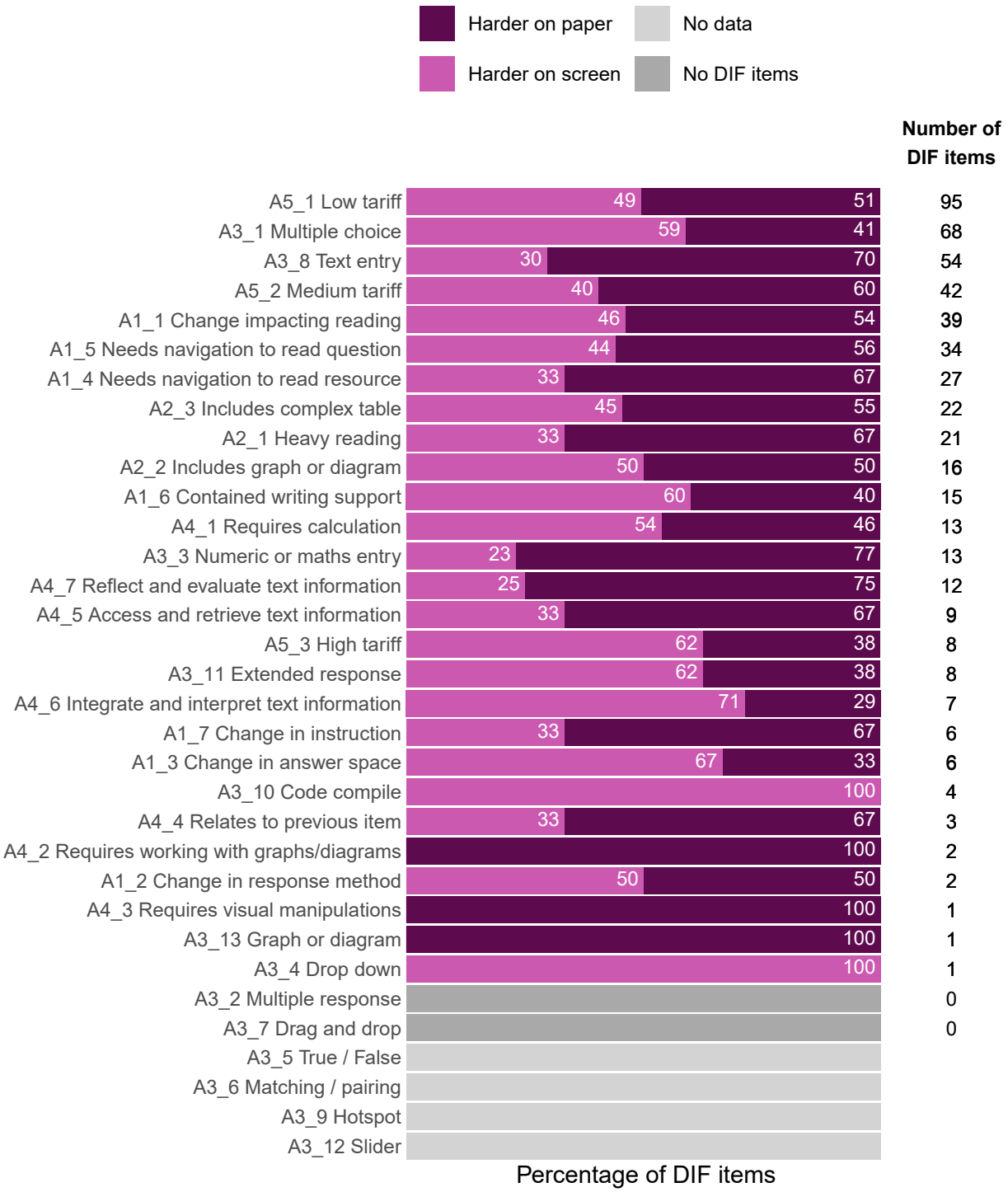

Figure 3: Percentage of Differential Item Functioning (DIF) items estimated to be harder on paper or on screen, and the total number of DIF items for each characteristic. Characteristics are ranked from top to bottom by the number of DIF items.

Conclusions

In this article, our first aim was to investigate whether the difficulty of an item remained broadly similar when tested on paper and on screen, using data from the Digital Mocks Service offered by Cambridge University Press & Assessment since 2023. This service allows candidates to take digital mock exams which are often based on past papers used in previous live exam sessions. Specifically, we

examined which items showed Differential Item Functioning (DIF). In this context, an item is said to exhibit DIF when candidates of similar ability show different performance depending on whether they answered it on screen or on paper. Even though the presence of DIF does not necessarily indicate the existence of a mode effect, it is an indication of unexpected functioning of the item.

Based on 795 items across 31 exams (from GCSE, Cambridge IGCSE and Cambridge International AS and A Level qualifications), our findings indicate that about one in five items (18 per cent) exhibit DIF. It is important to note that this finding cannot be generalised to other digital exams. This is because testing platforms as well as how the digital items are designed and presented to candidates on screen can vary in ways that influence their propensity to have a mode effect. It is challenging to gauge, based on existing literature, whether this percentage of DIF is worrying. The number of items with DIF vary greatly across studies: for example, results from Steedle et al. (2020, p. 10) indicated that 8 per cent of items in the analysed ACT tests exhibit DIF, while, in contrast, Gu et al. (2006) found that 38 per cent of the analysed GRE[8] Mathematics items were flagged for DIF.

The second aim of the research presented here was to investigate whether DIF items were more likely to be found in items with a specific characteristic. Our findings suggest that DIF, and hence potential mode effects, are unlikely to be systematically associated with a particular item characteristic. There were two exceptions. Firstly, items that require "numeric or maths entry" as answers showed a frequency of DIF that was higher (about three in 10 items had DIF) than other types of items, and with proportionally more items estimated to be harder on paper than on screen. There are a few possible reasons for the high frequency of DIF observed among these types of items. For instance, numeric entry fields on screen may encourage different formatting or lead learners to use a calculator differently. Secondly, the research also found that items requiring text entry as answers (not extended response) had a disproportionate number of DIF items found to be harder on paper than on screen. Students' preference for typing over handwriting could go some way towards explaining this finding (Whithaus et al., 2008). In both instances, further research might be needed to explore the underlying reasons for the patterns found.

Limitations

The results of the analyses carried out in this research should be treated as indicative rather than definitive due to the limitations outlined below.

Firstly, to classify items into a manageable number of categories, a simple item characteristics framework was proposed. As a result, many nuances were not captured. For example, the "complex table" characteristic subcategory may include items with tables of varying complexity and word count. Thus, we cannot entirely rule out the possibility that subsets of items within a characteristic category may be more prone to mode effects than others. Relatedly, it was challenging to capture the impact of each characteristic accurately within

8 GRE stands for Graduate Record Examinations. For more detail, see https://www.ets.org/gre.html.

this framework, especially those related to the differences in presentation and functionality across modes. This is because how the items were shown on the screen can vary across candidates, depending on several factors such as the screen size of the device used for the test and whether candidates have zoomed in. Future research can investigate how candidates interact with digital items (e.g., using eye-tracking or process data to validate behavioural assumptions), including how often they use the provided functionalities on the digital testing platform, and how that is linked to mode effects.

Secondly, the lack of data for certain item characteristics reduced our ability to draw definitive conclusions. We had no data on a handful of characteristics and very limited data on a few others, perhaps due to only a range of subjects being available in the Digital Mocks Service (e.g., Mathematics was not one of them). It was nonetheless reassuring that we did not find a strong relationship between item characteristics and DIF among those characteristics for which we had sufficient data.

Thirdly, disentangling the effect of individual item characteristics on performance across modes was difficult. For example, an item with a heavy reading demand would also involve scrolling and navigating, making it hard to determine which of the two factors may be driving any apparent mode effect. To isolate these effects, we recommend an in-depth study which would be either experimental, with controlled changes across test modes, or observational, with candidate behaviour observed during test-taking. Alternatively, we could explore the presence of DIF at item level by fitting regression models where the item characteristics are included as predictors.

Fourthly, DIF methods, particularly those based on IRT, require an external measure for ability so that item performance can be compared across groups with the same level of ability. When the total score on the test is used for this purpose (as has been done in this work due to not having data on an alternative measure), items with DIF can distort the total score, leading to potential inaccuracies in the estimation. However, in this work, as the number of items with DIF flagged in each test was relatively small and in different directions, the total score might not have been too affected by the item-level DIF.

Lastly, as taking digital mock exams is voluntary and schools may choose whether to offer them to all, some, or none of their candidates, differences in ability and motivation are possible between candidate groups across modes. Therefore, the detected DIF may partly be attributed to these group differences. The use of IRT to estimate item difficulty minimises this concern as it utilises candidates' performance across the entire exam to produce difficulty estimates that are comparable between groups. While this approach is more robust than simpler methods, such as comparing the percentage of correct responses in each item across modes, it does not entirely eliminate the possibility of bias.

References

Crisp, V., & Ireland, J. (2022). A structure for analysing features of digital assessments that may affect the constructs assessed. Cambridge University Press & Assessment.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

Fishbein, B. (2018). *Preserving 20 years of TIMSS trend measurements: Early stages in the transition to the eTIMSS assessment* [Unpublished doctoral dissertation]. Boston College.

Gu, L., Drake, S., & Wolfe, E. W. (2006). Differential item functioning of GRE mathematics items across computerized and paper-and-pencil testing media. *The Journal of Technology, Learning and Assessment, 5*(4).

Harrison, S., Kroehne, U., Goldhammer, F., Lüdtke, O., & Robitzsch, A. (2023). Comparing the score interpretation across modes in PISA: An investigation of how item facets affect difficulty. *Large-scale Assessments in Education, 11*(1), 8.

Hughes, S., Custodio, I., Sweiry, E., & Clesham, R. (2011, November 8–10). *Beyond multiple choice: Do e-assessment and mathematics add up?* [Paper presentation]. Annual conference of the Association for Educational Assessment – Europe (AEA-Europe), Belfast, Northern Ireland, UK.

Imran, M., Elaskari, S., & Albalawi, A. E. (2021). Comparative life cycle assessment of paper and computer-based exams. *International Journal of Thermal & Environmental Engineering*, *18*(2), 71–77.

Jerrim, J. (2016). PISA 2012: How do results for the paper and computer tests compare? *Assessment in Education: Principles, Policy & Practice*, *23*(4), 495–518.

Jerrim, J., Micklewright, J., Heine, J-H., Salzer, C., & McKeown, C. (2018). PISA 2015: How big is the 'mode effect' and what has been done about it? *Oxford Review of Education, 44*, 476–493.

Kuvalja, M. (2023, November 1–4). *Evaluation of the Cambridge International Digital Mock Exams Service* [Paper presentation]. Annual conference of the Association for Educational Assessment – Europe (AEA-Europe), Malta.

Li, D., Yi, Q., & Harris, D. (2017). Evidence for paper and online ACT® comparability: Spring 2014 and 2015 mode comparability studies. ACT.

Linacre, J. M. (2000). Comparing and choosing between "Partial Credit Models" (PCM) and "Rating Scale Models" (RSM). *Rasch Measurement Transactions, 14*(3), 768.

Linacre, J. M. (2023). Winsteps® Rasch measurement [Computer program]. Winsteps.com.

Mistry, S., & Kemmery, S. (2025, November 6–8). *Are schools ready for digital exams? – Creating a framework to evaluate digital readiness* [Paper presentation]. Annual conference of the Association for Educational Assessment – Europe (AEA-Europe), The Netherlands.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper & pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning and Assessment, 3*(6).

Rogo Assessment. (2025, March 28). *Can examinations ever be green? Sustainability in professional assessment*.

Steedle, J., Pashley, P., & Cho, Y. (2020). *Three studies of comparability between paper-based and computer-based testing for the ACT®.* ACT.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Lawrence Erlbaum Associates.

Vidal Rodeiro, C. L., Gill, T., & Hughes, S. (2024, November 7–9). *Using assessment and response times data to evaluate a digital mock exams service* [Paper presentation]. Annual conference of the Association for Educational Assessment – Europe (AEA-Europe), Cyprus.

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.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*(1), 1–28.

Predicting success in vocational education and training: Insights from literature

Annaliese Micallef Grimaud (Research Division)

Introduction

Cognitive and non-cognitive abilities, along with previous achievement, are well-established predictors of future examination attainment (e.g., Deary et al., 2007; Jencks et al., 1979; Poropat, 2009). These factors are frequently used to predict students' potential performance and success in future academic qualifications, such as GCSEs and A Levels[9] (Deary et al., 2007; Jellis, 2024). While there is a substantial amount of work on predictors of academic achievement, it seems that work on what predicts success in vocational qualifications is more limited.

In the English education system, vocational qualifications (VQs) are commonly defined as education and training programmes that adopt a predominantly skills-based approach and are designed to equip individuals with the skills and competencies required for specific occupations or professional sectors (DfE, n.d.; Ofqual, 2021). It is important to note that terminology relating to vocational and technical education is not internationally standardised and varies considerably across countries, policy contexts, and education systems. Vocational education and training has been described as a contested field, shaped by historical, political, and economic factors (McGrath, 2012). Terms such as vocational, technical, and skills-based education and qualifications may be used with different meaning and emphases across national contexts or may be used interchangeably. Comparative research further highlights that the diversity of Vocational Education and Training (VET) systems is reflected in the wide range of terms used internationally, and that similar terms may refer to different educational contexts, while different terms may be used to describe comparable practices (Chen et al., 2021). As a result, boundaries between vocational and technical qualifications and education are often blurred, and terminology is frequently used with overlapping or context-dependent meanings. In this article, the terms VQ and VET are used to refer to programmes that prioritise occupationally relevant skills and competencies, while acknowledging the diversity of structures and labels used internationally.

1 GCSEs and A Levels are academic qualifications taken by students in England. GCSEs refer to General Certificate of Secondary Education. These are exams used to measure academic attainment at ages 15 to 16. A Levels, short for Advanced Level exams, are higher-level qualifications typically taken after GCSEs by students at 17 to 18 years of age.

The structure and assessment of VQs tend to differ in emphasis from those of qualifications commonly categorised as academic, such as GCSEs and A Levels. While academic qualifications may include practical or internally assessed components, they typically place greater weight on subject-based knowledge assessed through written examinations. VQs, by contrast, often place a greater emphasis on applied assessment methods – such as coursework, portfolios, practical demonstrations, and in some contexts, work-based assessments – aligned with occupational or industry-specific competencies (Vitello & Williamson, 2017). As has been noted in assessment research, the distinction between academic and vocational qualifications lies less in the type of assessment instruments used, and more in their relative balance, intended purpose, and underlying assessment paradigm (Baird et al., 2024).

In England, it has been reported that the uptake of VQs at both secondary and post-16 levels has increased (Mitton & Hensby, 2024; Vidal Rodeiro & Vitello, 2020). This growth underscores the need for a clearer understanding of which factors may predict success in these pathways, particularly given their distinct assessment structures and educational purposes.

Existing literature reviews and empirical studies seem to have predominantly focused on academic achievement. For instance, Poropat (2009) conducted a meta-analysis investigating the relationship between personality traits and academic performance, but the review focused solely on academic qualifications and did not include vocational qualifications. Similarly, Wyness et al. (2023) explored how well previous grades can predict future grades in A Levels in the UK. However, they purposely excluded vocational qualifications from their analyses. Another study (Molnár & Kocsis, 2023) analysed cognitive and non-cognitive predictors of academic success of students at one of the highest-ranked universities in Hungary; however, the analysis treated the data as a whole and did not report information about courses, fields of study, or subject areas. As a result, it is unclear whether vocational or professionally oriented pathways were represented in the sample.

Some studies have included non-academic pathways, such as vocational and technical education, qualifications, or training, within broader analyses of educational attainment. However, in such cases, vocational pathways are typically studied alongside academic qualifications and are not always separated in the analyses. Moreover, some studies do not report subject areas, fields of study, or specific courses for either vocational or academic qualifications, limiting insight into how predictors may operate across different educational contexts. For example, Conlon (2005) explored what cognitive and non-cognitive elements may predict future academic and vocational qualification attainment in the UK. However, the study did not differentiate between pathways or provide subject-level information, limiting conclusions specific to vocational qualifications. Similarly, Williams et al. (2022) systematically reviewed 90 studies to determine which factors from middle childhood may predict academic attainment at 15 to 17 years of age in the UK. Although National Vocational Qualifications (NVQs)[1]

1 National Vocational Qualifications were competence-based qualifications in England that assessed occupational skills through practical, outcomes-based assessment, which were phased out during the 2010s as part of wider reforms to the vocational qualification system (Winch, 2021).

were included in the review together with GCSEs and other examinations, the different types of qualifications and subjects were not differentiated, preventing vocational- or subject-specific conclusions.

In addition to predicting subsequent attainment outcomes, it is interesting to also consider how individual characteristics may be associated with the educational pathways that students enter. Previous research has highlighted how factors beyond achievement play a role in these decisions. For example, socioeconomic background and family expectations have been associated with whether students pursue academic or vocational routes (Erikson & Jonsson, 1996; Esser, 2016). Other studies have proposed that aspects of students' personality traits may also contribute to the educational route chosen (Usslepp et al., 2020). These patterns suggest that certain traits and abilities may provide insight into which students choose which pathways, and thus into the typical profiles of those entering vocational education.

Understanding what predicts success in vocational education is important for shaping effective educational policy and providing guidance to students pursuing non-academic pathways. It can also help identify whether particular groups of learners are more or less likely to access, succeed in, or leave vocational pathways. This article presents findings from a scoping review of the literature on predictors of success in vocational education, training, and qualifications. While a wide range of factors may influence success in vocational education (e.g., socioeconomic status, demographic characteristics), the present review focuses on three broad sets of predictors – prior attainment, cognitive abilities, and non-cognitive abilities. This focus reflects a particular interest in factors that tend to be commonly measured through standardised assessments or structured instruments. The review was originally conceived with a particular focus on what factors predict success in vocational education in the UK, and although its scope was subsequently broadened to include international evidence in order to capture a wider range of relevant research, it retains a strong interest in drawing out implications for vocational education in the UK. The review considers both the association of these predictors with outcomes in vocational education and their potential links with the likelihood of students entering particular educational routes (i.e., their role in pathway selection).

Method

Since the aim was to identify the extent of existing research on what predicts success in vocational education in the UK, a scoping review was conducted. Given the limited and fragmented nature of research in this area – particularly within the UK context – a scoping review method was appropriate to explore the breadth of available evidence without restricting the analysis to a narrow set of study designs or outcomes. Following the guidelines outlined by the Joanna Briggs Institute (Peters et al., 2015) and Tricco et al. (2016), a research protocol was devised, which specified the search terms to be used, the databases to be searched, and the inclusion and exclusion criteria. The study selection process entailed title and abstract screening, followed by full-text screening for studies that seemed relevant. Articles included in the review were then coded in MAXQDA

Plus 24 (Release 24.7.0) (VERBI Software, 2025) based on predictor types and qualification subjects. The whole process of the scoping review was conducted by one researcher.

Search criteria

Searches for peer-reviewed articles and book chapters were conducted in Scopus, Springer, JSTOR, ERIC, and ProQuest databases, as well as Cambridge University's online library database. Searches were also run in Google and Google Scholar. Search terms included "predictive validity", "prior attainment", "predictors" and "vocational qualifications" used in combinations (e.g., predictive validity AND vocational qualifications).[2] Quantitative studies, particularly correlation studies and regression studies, meta-analyses, and literature reviews were also searched for by combining these words with the previously mentioned search terms (e.g., meta analysis AND predictive validity AND vocational qualifications). The search was limited to articles written in the English language.

Inclusion and exclusion criteria

Due to the special interest in the UK education system, the search was initially restricted to UK-based studies that investigated predictors of success in vocational qualifications. Additionally, only studies published from 2017 were initially included, reflecting policy reforms that reshaped vocational education pathways during this period.[3] However, due to the small number of studies identified under the original inclusion criteria, both the publication year restriction and initial UK-only focus were removed in order to capture a broader and more informative evidence base. As a result, studies from a wide range of time periods and international contexts were included. It is acknowledged that findings from earlier studies may not fully reflect current educational contexts due to changes in education systems over time, and that differences between national education systems may limit the direct comparability or transferability of findings across countries. Accordingly, findings from older studies are interpreted cautiously and used primarily to identify broad patterns in predictors of vocational success, rather than to draw direct conclusions about current vocational qualifications.

The search was conducted using terms such as "vocational qualifications" (see the Appendix); however, studies using related terminology (e.g., vocational education, training, technical programmes, apprenticeships), although not explicitly used as search terms, were nevertheless identified through the search and included. This reflects the wide variation in terminology used internationally to describe vocational education and training. In addition, some studies identified in the search reported insights on how certain factors (such as personality traits) may influence students' choice of educational pathway. Although pathway choice was not the primary focus of the review, these studies were retained because they may provide relevant contextual information about the profiles of students who enter a vocational pathway.

2 The full list can be found in the Appendix.

3 Information about policy reforms can be viewed here: https://www.gov.uk/government/publications/technical-and-further-education-bill-factsheet

The articles found were included in the screening if they were:

- Empirical papers (such as correlation or regression analyses) linking variables (prior attainment, cognitive abilities, non-cognitive abilities) to predicting success in VQs (or similar)
- Meta-analyses or literature reviews linking variables (prior attainment, cognitive abilities, non-cognitive abilities) to predicting success specifically in VQs (or similar), or in VQs (or similar) alongside other types of qualifications
- Studies that link prior attainment, cognitive abilities, and / or non-cognitive abilities to the likelihood of students choosing a vocational track

Screening

Search results were initially screened by reviewing titles and abstracts to assess whether they met the inclusion criteria specified in the research protocol. Of 113 potentially relevant articles screened in this way, 63 were excluded. The remaining articles were obtained and read in full, which led to a further 32 articles being excluded as they did not meet the inclusion criteria.

A final number of 18 articles were included in the review. A summary of the search and screening process is presented in Figure 1, in the form of a PRISMA 2020 (Page et al., 2021) flow diagram. A table listing the studies included in the review together with their key characteristics is presented in the Supplementary Materials (Table S1).

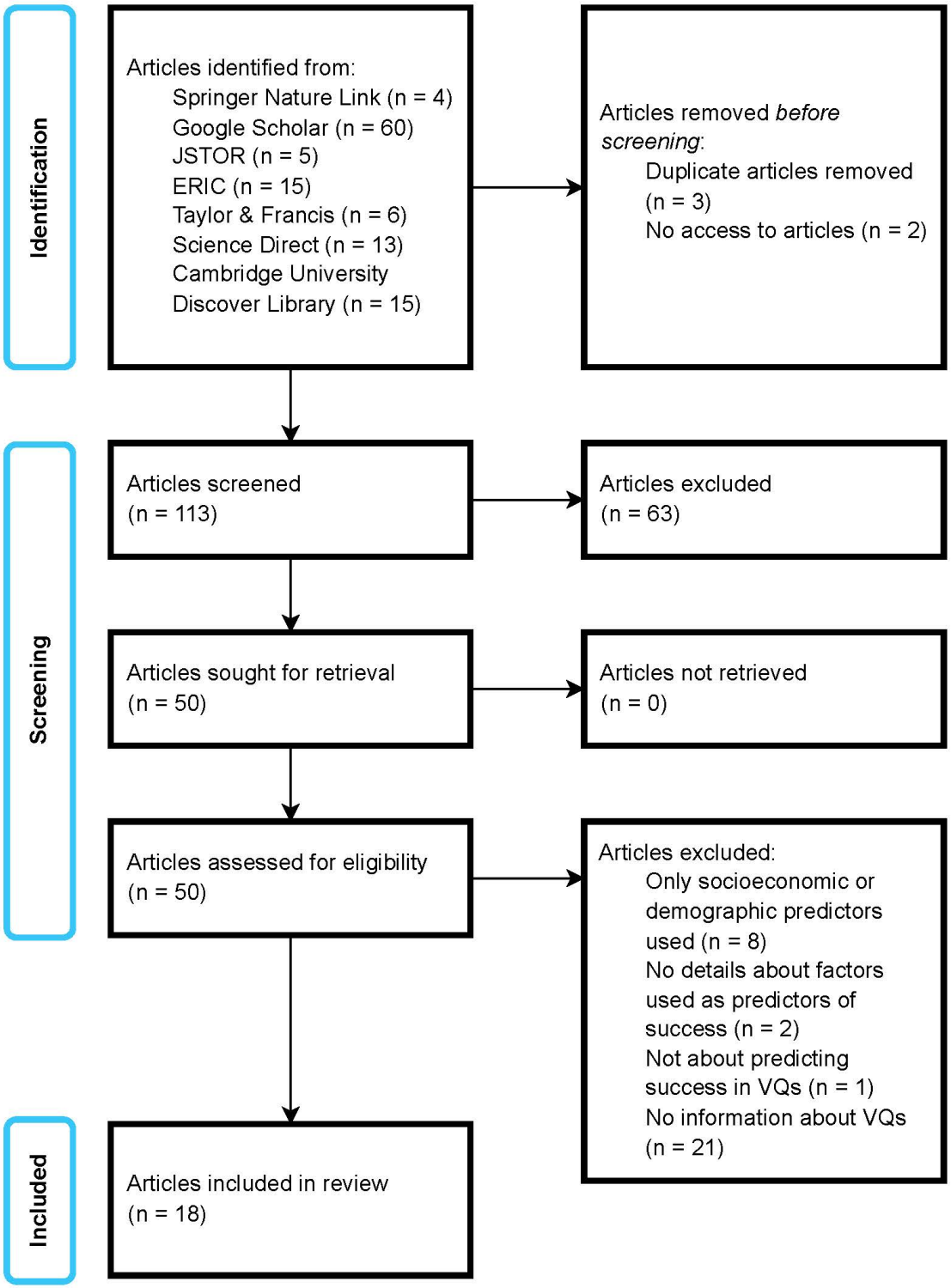

Figure 1: PRISMA 2020 flow diagram of article selection (Page et al., 2021)

Findings

This section reports findings regarding predictors of success in vocational qualifications and training, followed by findings relating to factors influencing students' choice of educational route.

Types of predictors

The findings for factors predicting success are organised by type of predictor. First, studies relating to prior academic achievement are presented, followed by ones about cognitive abilities, and then non-cognitive abilities. Some studies focused on one type of predictor (e.g., prior attainment), whereas others investigated combinations of multiple predictors (e.g., cognitive abilities and personality traits).

Prior academic achievement as a predictor

Some studies have used prior attainment in academic qualifications (Klapproth & Schaltz, 2014; Vidal Rodeiro & Vitello, 2020) or prior vocational qualifications (Barnard & Dixon, 1998) as potential predictors of future attainment. These include scores from previous examinations, such as GCSE grades (Vidal Rodeiro & Vitello, 2020).

Evidence has consistently shown that prior attainment predicts outcomes in vocational programmes, albeit more weakly than in academic routes. In a meta-analysis which included 63 studies from Germany, Baron-Boldt et al. (1988), as reported by Klapproth and Schaltz (2014), found that average school grades were slightly more strongly correlated with success in academic qualifications (r = 0.46) than in vocational ones (r = 0.37).

A study based on England's General National Vocational Qualifications (GNVQs) – VQs offered in the UK which were discontinued, with last awards in 2007 – reported a similarly weak correlation (around r = 0.30 and explaining about 8–9 per cent of the variance) between Advanced GNVQ grades (Level 3 qualifications, broadly equivalent to two A Levels) and prior average GCSE score (Barnard & Dixon, 1998). By contrast, GCSE scores were reported as typically accounting for up to 50 per cent of variance for A Level grades. Barnard and Dixon's analyses also showed that prior success on lower-level vocational awards predicted subsequent vocational success, but the average GCSE score was determined to be the best predictor of GNVQ attainment from the variables investigated.

Another study based in the UK by Schofield and Dismore (2010) also found prior attainment as the strongest predictor from the variables they investigated. The study investigated whether entry grades (UCAS points),[4] age, gender, the type of previous academic qualification, the timing of application, and the duration of time elapsed since a student's last qualification could serve as predictors for retention and achievement of students in Foundation Degrees (FDs) and

4 UCAS points refer to the numerical values assigned to certain UK qualifications under the UCAS Tariff, which is a standardised system designed to support comparison of applicants' attainment for entry to higher education across different qualification types (UCAS).

Higher National Diplomas (HNDs), both of which are vocationally oriented higher-education qualifications in the UK. The study reported that entry grades were the strongest predictor for both retention and achievement of students. Students who progressed to their second year of study in FD and HND programmes had 20 per cent higher UCAS points on entry (2010, p. 217).

Large-scale analyses by Ofqual (2024) presented similar findings. In analyses predicting qualification outcomes, prior attainment at Key Stage 4 was consistently found to be the strongest predictor across a range of qualification types, including vocational and technical qualifications, GCSEs, and A Levels. Other variables examined in the analyses included region, special education needs and disability status, first language, and region; however, these factors showed weaker associations with qualification outcomes than prior attainment.

Patterns suggesting that prior attainment is an important, but not universally dominant, predictor of success are also evident in earlier studies conducted in the United States (US). For instance, in a review of 38 studies, Prediger et al. (1968) synthesised findings about nine variable categories investigated as predictors of success in 11 high school level vocational education programmes in the US. They summarised over 1 200 correlations between variables and success of students in the following high school level vocational education programme categories: Auto Mechanics, Carpentry, Drafting, Electricity, Machine Shop, Industrial Arts, Business Education, Bookkeeping, Shorthand, Typing, and Home Economics. To categorise the diverse predictors found across studies, Prediger et al. adapted a classification system previously used by Ghiselli (1966) for aptitude tests. They also introduced additional categories for specific purpose aptitude tests, past grades, and achievement tests, as a practical solution to group the varied measures encountered in the literature. The measures that they used for prior attainment were Grade Point Average (GPA)[5] and achievement test data (ACH) – a variable that referred to scores from tests typically assessing areas such as reading, spelling, or English grammar. The GPA and ACH variables showed the strongest correlations with Shorthand and Business Education courses (r = 0.44-0.56). However, Prediger et al. noted that the strength of median correlations between predictors and success in the courses varied considerably, with no specific variable being the strongest predictor across all courses.

Similar conclusions were reported in another early review by Stock and Pratzner (1969), which synthesised findings from a wide range of US studies on predictors of success in vocational education. In that review, previous attainment (measured via GPA) was reported as the strongest predictor for success in Nursing and Technical (e.g., electrical industrial drafting and design, civil engineering) programmes, whereas other variables from among those reviewed (e.g., verbal and reading abilities, spatial abilities, etc.) had stronger correlations for other courses. Together, these early studies illustrate that while prior attainment is often a powerful predictor, its strength – and that of other predictors – varies by subject and programme.

5 The Grade Point Average (GPA) is a numerical measure of overall academic achievement used in the US. GPA is calculated by assigning point values to letter grades and averaging these across completed courses, typically on a 4.0 scale, with 4.0 being the best.

Cognitive abilities as predictors

A number of cognitive abilities have been tested as potential predictors of success in VQs or VET. The most frequently investigated cognitive abilities include reading and verbal abilities, mathematical and numerical skills, non-verbal reasoning, and spatial abilities. These tend to be packaged as subcomponents of general mental ability (GMA) or similar intelligence tests. Other less frequently reported cognitive abilities assessed include divergent thinking and practical-scientific competence (Diedrich et al., 2018), but their appearance in the literature reviewed was sporadic.

General intelligence

In a large-scale UK investigation, Deary et al. (2007) explored the relationship between general intelligence (denoted as *g*) measured at age 11 using the Cognitive Abilities Test (CAT; comprising verbal, quantitative, and non-verbal reasoning batteries),[6] and students' subsequent examination performance in 25 GCSE and GNVQ subjects. These subjects were grouped into Arts and Humanities, Science, Social Sciences, and Practical subjects. While several subjects included in the analysis (e.g., English, Mathematics, Double Science, French, and Geography) were GCSEs, the study did not specify which subjects were taken as GNVQs. As such, the findings cannot be used to isolate predictors of vocational qualification success but are included here to illustrate how the predictive strength of general intelligence varies across subject domains, including applied or performance-based subjects that overlap conceptually with vocational education. In this section, findings relating to general intelligence are presented; findings relating to individual abilities using one of the batteries mentioned are reported in subsequent sections.

Deary et al. found that intelligence strongly correlated with overall educational achievement (r = 0.81). However, the strength of correlations varied between the subject groups. The correlations between CAT *g* and subject group are presented in Table 1. Correlations between subject group and the verbal battery of CAT (referred to as CAT residual verbal) are also available in Table 1, however, these are discussed in the next section. Correlations were highest for the Sciences group (r = 0.46-0.77), particularly for Mathematics (r = 0.77). Associations were lower, although still moderate, for the Practical subjects such as Physical Education, Music, and Art and Design (r = 0.43-0.55). These patterns highlight domain-specific variation in the predictive validity of general intelligence, with applied subject areas showing slightly weaker associations compared to more traditionally academic subjects.

6 The authors refer to the variable measuring general intelligence (*g*) using the CAT as CAT *g*.

Table 1: Correlations between subject groups, CAT *g*, and CAT residual verbal (Deary et al., 2007)[7]

Subject group	Correlation with CAT *g*	Correlation with CAT residual verbal
Arts and Humanities: English, English Literature, Drama, Religious Education, French, German, Spanish	*r* = 0.47-0.67	*r* = 0.14-0.22
Science: Mathematics, Double Science, Single Science, Physics, Chemistry, Biology	*r* = 0.46-0.77	*r* = 0.00-0.14
Social Sciences: Geography, History, Business, Information Technology, Information Technology short course	*r* = 0.47-0.65	*r* = 0.07-0.18
Practical: Art and Design, Music, Physical Education, Design and Technology – Food, Design and Technology – Graphics, Design and Technology – Resistant Materials, Design and Technology – Textiles	*r* = 0.43-0.55	*r* = 0.06-0.16

Subject-dependent correlations were also observed in a review by O'Rourke (1938) who reviewed over 500 studies examining aptitude measures across multiple domains. O'Rourke reported particularly strong correlations between intelligence and performance in medical and health-related subjects. For example, the Otis Intelligence Test was found to be a strong predictor of first year mean scores of dental school students ($r = 0.55$) and an even stronger relationship was reported between Intelligence Quotient (IQ) and nurse examination scores ($r = 0.72$). However, different relationships were noted for general intelligence and music. For example, O'Rourke reported that a study by Kwalwasser (1936) found little association between general intelligence and musical intelligence, while Farnsworth (1935) found that music capacity tests served as better predictors than intelligence tests for courses involving tonal perception and musical performance. It should be noted that O'Rourke's meta-analysis was conducted in 1938, and that definitions and measures of intelligence have evolved considerably over time. As a result, these historical differences may limit the extent to which the findings are relevant to current educational contexts (see limitations section for further discussion).

A similar pattern appears in more recent research on vocational training programmes. Bertua et al. (2005) carried out a meta-analysis examining the predictive validity of cognitive ability tests for training success across a range of occupations in the UK. The tests included General Mental Ability (GMA), as well as verbal, numerical, perceptual, and spatial ability tests. Occupations were

7 See next section for commentary about the findings for CAT residual verbal. Table reprinted from Intelligence, Volume 35, Issue 1, Ian J. Deary, Steve Strand, Pauline Smith, & Cres Fernandes, Intelligence and educational achievement, Page 9, Copyright (2007), with permission from Elsevier. License number: 6145380716066.

grouped into the following categories: clerical jobs, engineers, professionals (e.g., nurses, student nurses, dental students), drivers, operators (e.g., radar and naval operators, naval coders, cipher operators, air traffic controllers), skilled workers (e.g., electricians, radio and electrical mechanics), and a miscellaneous / mixed category (e.g., cooks, military clerks, drivers, police).

GMA tests showed high predictive validity for training success across all occupational groups. The strongest predictive validity coefficient was observed for the engineering group (ρ = 0.64), while the weakest was observed for the driver group (ρ = 0.47). Table 2 summarises the predictive validity coefficients for GMA tests across the different occupational groups. The meta-analysis did not offer a similar breakdown for the other ability tests analysed.

Table 2: Predictive validity coefficients (ρ) for GMA tests predicting training success (Bertua et al., 2005)[8]

Occupation group	GMA test predictive validity coefficient (ρ)
Engineering	0.64
Professional	0.59
Clerical	0.55
Skilled	0.55
Miscellaneous / mixed	0.55
Operator	0.54
Driver	0.47

Overall, these findings suggest that general intelligence is a strong predictor of educational performance and training success across a wide range of vocational domains. However, the predictive strength varies by subject and occupational context. This variation provides a rationale for examining specific cognitive abilities to understand more precisely how different aspects of ability relate to success in vocational education and training.

Verbal, numerical, and non-verbal abilities

Verbal, numerical, and non-verbal abilities tend to be assessed as part of general intelligence, but some studies have examined how the individual components correlate with performance in different subjects and vocational domains.

Verbal ability as a predictor of future attainment has been investigated in a number of settings. Deary et al. (2007) investigated the association between verbal ability – measured using the CAT residual verbal component of the CAT general intelligence measure – and students' performance in GCSEs and GNVQs. As with the analyses of general intelligence, the study did not distinguish between academic and vocational qualification types within subject areas, but

8 Reprinted from Journal of Occupational and Organizational Psychology, Volume 78, Issue 3, Cristina Bertua, Neil Anderson, and Jesús F. Salgado, The predictive validity of cognitive ability tests: A UK meta-analysis, Page 23, Copyright (2011), with permission from John Wiley and Sons. License number: 6147610492044.

provides insight into how verbal ability related to attainment across different subject groups, including applied and performance-based subjects. Deary et al. found that correlations between performance and CAT residual verbal ability were generally small across domains, and notably weaker than those observed for general intelligence (see Table 1). Correlations ranged from r = 0.00 for Mathematics to r = 0.22 for English, which was the strongest correlation with this measure.

In their multi-study review, Prediger et al. (1968) adopted and defined "V-INTEL" as a predictor category specifically designed to encompass tests of verbal intelligence and academic aptitude. Their aggregated findings showed that V-INTEL yielded the strongest correlations for Business Education, Bookkeeping, and Shorthand courses (r = 0.40-0.44), and the weakest correlations for Carpentry and Electricity (r = 0.14-0.18). A similar subject-specific pattern was reported by Stock and Pratzner (1969), who examined predictors across multiple vocational programmes. In their review, verbal and reading abilities (r = 0.44) and English grades (r = 0.44) were among the strongest predictors of success in Business Education programmes, whereas reading and verbal abilities showed weak correlations with success in Auto Mechanics. Taken together, these findings suggest that verbal abilities may predict success in some vocational programmes, but that the strength of prediction varies across vocational subjects.

Diedrich et al. (2018) employed hierarchical regression analyses to examine the predictive validity of verbal, numerical, and spatial abilities, among other cognitive and non-cognitive abilities, for the grades and satisfaction of students in Austrian higher vocational schools and apprenticeships. Their sample covered occupations grouped in Food, Tech, People, Paper, and Craft branches. Table 3 presents the occupations within each branch.

Table 3: The occupations within each branch presented in Diedrich et al. (2018)

Branch	Occupations
Food	Agricultural school, baker, cook, gastronomy
Tech	Engineering school, metal technician, electrical engineering, Information Technology
People	Nursing, BAKIP (kindergarten teacher training), hairdresser, florist
Paper	Academic secondary school, commercial academy (HAK), office, sales
Craft	Joiner, bricklayer, installations (plumber)

Notes. Academic secondary school refers to the traditional high school that leads to academic studies. Commercial academy (HAK) is a type of vocational high school in Austria that specialises in commercial and economic education.

The predictive validity strength of abilities varied across the five occupation branches. Verbal ability significantly predicted grades in the Food (β = 0.23), Craft (β = 0.23), and Tech (β = 0.21) branches, and spatial ability was also a significant predictor (β = 0.17) in the Food branch. In the Paper branch, numerical ability initially emerged as the only significant predictor of grades (β = 0.29). However,

after facets of Conscientiousness[9] (specifically ambition, β = 0.34) and vocational interests (specifically science interest, β = 0.20) were added to the model, verbal ability also became a significant predictor of GPA (β = 0.24). This suggests a more nuanced role for verbal ability, contributing to success in some vocational branches directly and in others once personality and interests were taken into account.

Findings for numerical abilities show similarly strong but domain-specific patterns. Prediger et al. (1968) reported that the category grouping arithmetic reasoning and computation variables (labelled ARITH) was one of the predictors with the largest variability in correlations, ranging from a weak correlation (r = 0.06) with Carpentry courses, to a moderately strong correlation (r = 0.48) with Business Education courses. Stock and Pratzner (1969) similarly found that arithmetic ability was the best predictor for the Business Education programme (r = 0.48), followed by verbal and reading abilities (r = 0.44) and English grades (r = 0.44). In the same review, arithmetic ability also showed a strong correlation with success in the Drafting programmes (r = 0.53), which referred to industrial drafting and design technology courses offered at the time. In Diedrich et al.'s (2018) study, numerical ability significantly predicted grades in the Paper branch (β = 0.29). This branch included academic secondary school, commercial academy, office, and sales occupations, suggesting that numerical aptitude was important across the different programmes within the branch. In their meta-analysis on how well cognitive ability tests could predict training success across a number of occupations in the UK, Bertua et al. (2005) reported that numerical ability tests served as the strongest predictor for training success (ρ = 0.54) across all the occupation groups under investigation (Engineering, Professional, Clerical, Skills, Operator, Driver, Miscellaneous). This was followed by perceptual ability tests (ρ = 0.50), verbal ability tests (ρ = 0.49), and spatial ability tests (ρ = 0.42), suggesting a modest advantage for numerical over the other specific abilities investigated in predicting training success.

Several non-verbal cognitive abilities have also been investigated, including spatial ability, mechanical reasoning, and perceptual speed and accuracy. Prediger et al. (1968) reported correlations between five non-verbal cognitive ability categories in relation to different programmes: non-verbal intelligence and abstract reasoning (NV-INTEL), spatial aptitude (SPACE), understanding of mechanical concepts (MECH), and perceptual speed and accuracy (PERCEPT). MECH abilities showed the strongest correlation with success in Machine Shop courses[10] (r = 0.44). SPACE abilities served as the strongest predictors for Drafting programmes[11] (r = 0.42) but correlated only weakly to moderately with the other courses, ranging from r = 0.15 to 0.36. The NV-INTEL group had its

9 Conscientiousness is one of the personality traits of the Big Five framework (McCrae & John, 1992), and it represents the extent of someone's organisation, persistence, and goal-directed behaviour. More details are given in the Big Five personality traits section of this article.

10 "Machine Shop" courses refer to a vocational programme category that includes courses named "Machine Shop", "Machine", and "Machinist". These courses typically focused on teaching students practical skills in metalworking and machining.

11 Prediger et al. (1968) also included courses with titles such as "Mechanical Drafting" and "Mechanical Drawing" in their Drafting category (p. 134).

strongest correlations with Home Economics and Machine Shop (r = 0.43-0.46), and its weakest correlations with Bookkeeping (r = 0.11). The PERCEPT group was consistently among the weakest predictors across most vocational courses.

Stock and Pratzner (1969) reported similar findings for spatial and mechanical abilities, noting that the strongest correlations often involved composites of multiple aptitude variables. For example, a combination of abilities and constructs derived from the Differential Aptitude Tests (DAT),[12] which assess abilities such as mechanical reasoning and spatial reasoning, were found to best predict success in Auto Mechanics and Machine Shop programmes (DAT composite r = 0.66-0.76). Similarly, the strongest predictor for the Drafting programme was a composite of the Flanagan Aptitude Classification Tests score (FACT, assessing spatial and mechanical abilities) and ACT (a university admissions test assessing academic performance) score (r = 0.67). In their study assessing prior attainment (using the average GCSE score) as a potential predictor of GNVQ grades, Barnard and Dixon (1998) also discovered that a spatial ability test predicted success in GNVQ Art and Design with a similar level of predictive strength to prior attainment. This suggests that domain-relevant non-verbal abilities can match the predictive power of prior academic attainment in certain vocational areas.

Overall, evidence across these studies suggests that specific verbal, numerical, and non-verbal abilities may all contribute to predicting success in vocational education and training, but their relative importance varies markedly by subject, occupational branch, and the configuration of other variables (such as personality traits and interests) included in the analytical models.

Non-cognitive abilities as predictors

A number of studies have investigated non-cognitive abilities as predictors of success in vocational education. These included personality traits – particularly the Big Five personality traits (McCrae & John, 1992) – vocational interests, social-emotional competencies, motivation, and behavioural engagement. Studies that investigated non-cognitive abilities as predictors of future success tended to do so alongside cognitive abilities and / or indicators of prior academic attainment.

Big Five personality traits

Personality traits under investigation were typically operationalised using the Big Five framework by McCrae and John (1992), which distinguishes five broad trait domains: Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness. Neuroticism reflects a tendency towards emotional volatility and negative affect; sometimes Emotional Stability is measured instead, which is viewed as the opposite of Neuroticism. Extraversion measures sociability and energetic engagement with others, Openness to Experience is a measure of how curious and open to new experiences someone is, Agreeableness measures someone's cooperativeness and concern for others, and Conscientiousness represents the extent of someone's organisation, persistence, and goal-directed behaviour.

12 DAT is a battery of tests which measures abilities such as abstract reasoning, mechanical reasoning, and numerical ability. More information on the development of DAT and what it measures can be found here: https://dictionary.apa.org/differential-aptitude-tests

When it comes to the Big Five traits, Conscientiousness has consistently emerged as a strong predictor of success in studies on Vocational Education and Training (VET). In a German study about VET students, Nießen et al. (2020) found Conscientiousness to be the most robust predictor – compared to the rest of the Big Five traits, cognitive ability, and socioeconomic status (SES) – for obtaining and starting a VET position, final VET grades, and students' satisfaction with their VET experience. They did note that Conscientiousness was less effective in predicting dropout intentions. Extraversion also emerged as a significant predictor of the final VET grade, with higher levels of Extraversion associated with higher VET grades. Furthermore, these effects were stronger predictors than the cognitive ability measure used.

A similar trend of Conscientiousness being a strong predictor seems to hold in studies that investigated an even bigger number of variables simultaneously. Diedrich et al. (2018) also assessed the predictive validity of personality traits using the Big Five personality factors, alongside cognitive abilities (intelligence measured by verbal, numerical, spatial abilities) and other non-cognitive abilities and interests, on the grades and satisfaction of students in higher vocational schools and apprenticeships in Food, Tech, People, Paper, and Craft occupational branches. Across vocational branches, Conscientiousness – together with verbal and numerical intelligence – was consistently among the strongest predictors of success. Conscientiousness particularly emerged as a significant predictor of success in the Food, Paper, and Craft branches, contributing notably to the explained variance in grades. The remaining Big Five traits – Extraversion, Emotional Stability, Openness, and Agreeableness – did not show significant predictive value for success.

Vocational interests

Beyond cognitive and personality variables, Diedrich et al. (2018) also reported vocational interests (defined as preferences for specific activities or environments) and non-interests of students (understood as a significant lack of interest or even an aversion to particular vocational domains) as playing a role, albeit a small one, in predicting success and / or satisfaction in specific vocational branches. They explored students' vocational interests in 14 different domains (e.g., Craft, Law, Art, Information Technology, Gastronomy, etc.), and reported that, a non-interest in Information Technology, together with verbal ability and dutifulness (from Conscientiousness) explained 37 per cent of variance in grades in the Food branch. Numerical intelligence, ambition (from Conscientiousness), and a vocational interest in science explained 37 per cent of variance in grade success for the Paper branch. A non-interest in sales contributed to the prediction of grades in the Crafts branch, together with intelligence (specifically verbal abilities) and Conscientiousness (self-discipline facet), while a non-interest in social occupations negatively correlated with success in the Tech branch. Based on their findings, the authors proposed that the individuals' non-interests in particular areas may be as important as their interests in determining their future success.

Stock and Pratzner (1969) also reported that measures of vocational interest could predict success in vocational programmes, but its predictive power tended to be weak or inconsistent. The authors suggested that interest measures may be

relatively more effective for identifying which students were more likely to choose certain vocational programmes, rather than for predicting their success in the programmes. Samuelson (1958) investigated the relationship between interest profiles measured using the Kuder Preference Record (KPR) – a standardised vocational interest inventory comprising multiple scales reflecting preferences for 10 activity types[13] – and instructor judgements of student performance in trade school programmes. While two specific KPR scales, Persuasive and Literary (reflecting interests in social influence and verbal, language-based activities, respectively), were negatively correlated with performance in trade-based programmes, the overall predictive value of the KPR could explain 9 per cent of the variance in student success. These findings suggest that vocational interest alone may be insufficient to predict vocational success but may be more effective when used as a supplement to other predictors.

Volodina et al.'s (2015) findings also support this notion; vocational interests were reported as predicting success in the first phase of vocational apprenticeships for technicians (e.g., industrial and electrical technicians) and industrial clerks more reliably than cognitive ability alone.

Social-emotional competencies

In their investigation of predictors of success and satisfaction across five vocational branches, Diedrich et al. (2018) also examined social-emotional competence, creative potential, practical-scientific competence, and personality and motivation. Social-emotional competence refers to an individual's ability to understand and manage their own emotions, as well as to perceive and respond to the emotions of others. In their study, Diedrich et al. measured social-emotional competence by getting participants to choose how they would typically react to various emotionally charged scenarios. They found that social-emotional competence provided a small, yet significant, contribution to predicting success in the People branch (vocational track for nurses, kindergarten teachers, hairdressers, and florists) and students' satisfaction in the People and Food branches.

A longitudinal study of 219 vocational training students in welding and fitting, administration, accounting, and professional cooking programmes in Canada (Gilbert et al., 2024) similarly investigated emotional competence as a predictor of educational success. The study reported that interpersonal emotional competence positively predicted self-reported educational success, but showed a very small and non-significant correlation with the students' actual performance outcomes as measured by the ratio of successful assessment results to total assessment attempts (r = 0.003).

Motivation and behavioural engagement

In their study on GNVQs, Barnard and Dixon (1998) reported that motivation, attitudes, and aspirations were associated with GNVQ achievement, albeit modestly. In their 2021 study on training programmes in Quebec, Canada, Dubeau et al. found that motivation – measured through expectancies of success

13 The 10 scales are: outdoor, mechanical, computational, scientific, persuasive, artistic, literary, musical, social service, and clerical. A description of KPR can be found here: https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft52224-000

(students' perception of their own ability to succeed in a school subject or task) – significantly predicted grades for students in Vocational Training programmes, which correspond to International Standard Classification of Education[14] (ISCED) Level 3. However, this was not the case for students in Technical Training programmes, which correspond to ISCED Level 4, where neither motivation nor the other variables investigated could successfully predict grades in their programme. In an earlier study by Dubeau et al. (2017), motivation also emerged as a significant predictor of achievement for vocational training students, with students that exhibited higher self-efficacy and task value generally showing greater success and persistence in their programmes. The authors also measured students' behavioural engagement with the course (such as attending classes regularly) and found that together with motivation, it was also a significant predictor of success. It is important to note that prior academic achievement was also a measure in Barnard and Dixon's (1998) and Dubeau et al.'s (2017) investigations and served as a significant predictor together with the others in both investigations. Furthermore, Barnard and Dixon retained that prior academic achievement was the best predictor from the variables they investigated.

Creative potential, practical-scientific competencies, and dexterity

In their extensive study, Diedrich et al. (2018) also investigated creative potential, assessed through divergent thinking tasks, and practical-scientific competencies[15] as potential predictors of success in vocational branches. Results found creative potential only predicted satisfaction in the Paper branch, while practical-scientific competencies did not have any predictive power in any of the branches.

In their summary of 38 studies, Prediger et al. (1968) found that manual dexterity consistently showed some of the weakest correlations as a predictor of success in vocational courses, a trend also observed for perceptual speed and accuracy. These findings suggest that these constructs are less central as predictors of vocational success than other variables investigated in the same studies.

Pathway and selection effects

In addition to predicting success in vocational education, some predictors are also associated with the educational routes that students select in the first place. An example of this is prior attainment. Vidal Rodeiro and Vitello (2020) found that students pursuing more vocational pathways in post-16 education in England tended to have lower attainment in previous academic grades, particularly in GCSE Mathematics and English. The authors note that this lower academic attainment may act as a barrier to more academic options and, in turn, may influence students to take more vocational pathways. Williamson and Carroll (2018) presented similar findings where students taking Cambridge Technicals[16]

14 https://ec.europa.eu/eurostat/statistics-explained/index.php?title=International_Standard_Classification_of_Education_(ISCED)

15 Practical-scientific competence refers to the ability to apply natural science principles to everyday problems.

16 Cambridge Technicals are vocational qualifications at Level 2 and Level 3 for students over 16 years of age. More information on Cambridge Technicals can be found here: https://www.ocr.org.uk/qualifications/cambridge-technicals/

tended to have lower prior attainment scores than the ones taking A Levels. This pattern may partly explain why studies have reported weaker correlations between prior academic attainment and vocational qualification success: if the attainment range of students on vocational pathways is lower and narrower, correlations with outcomes are likely to be reduced.

Differences between students' educational pathways appear to be influenced by other factors as well as prior attainment. In a four-year longitudinal study tracking 1 351 Finnish students, Vasalampi et al. (2023) found that verbal and reading abilities of students also differed between different school tracks. Students with higher PISA (Programme for International Student Assessment) reading scores (reading proficiency / verbal ability) at the end of compulsory school were more likely to choose an academic track than a vocational one in upper secondary education (school at 16 to 19 years of age). It is good to note that the PISA reading scores did not predict the course completion of students in either of the selected tracks. However, course non-completion was more common among students in the vocational track than in the academic track.

Personality traits and cognitive abilities likewise appear to differ by track. Brandt et al. (2020) reported that personality traits (Emotional Stability, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness) and cognitive ability, specifically fluid intelligence (described by the authors as a reasoning capacity demonstrated by the ability to infer logical rules from abstract patterns), of Grade 9 students in Germany differed based on school tracks – vocational, intermediate, and academic. Students in the vocational track had the highest mean scores for Conscientiousness and Agreeableness, while students in the academic track had the lowest. In contrast, students in the academic track had the highest mean scores for cognitive ability, Openness, and Extraversion, with vocational-track students having the lowest scores on these variables. Mean scores for the remaining traits (Emotional Stability and Agreeableness) showed only small differences across tracks.

Overall, these findings suggest that prior attainment, verbal and reading abilities, personality traits, and cognitive abilities are not only predictors of achievement, but are also linked to students' selection of academic versus vocational pathways. As a result, vocational cohorts may differ from academic ones in their profiles of attainment and non-cognitive traits, which may in turn influence the observed predictive validity of different variables in vocational education and training.

Limitations

A key limitation of this review concerns the conceptual and terminological ambiguity surrounding vocational qualifications and education more broadly. As discussed earlier, terms such as vocational, technical, skills-based education, and training, are used with varying meanings across countries and policy contexts. This lack of standardisation means that studies included in the review often refer to courses and qualifications that differ in structure, purpose, and assessment. A related limitation arises from the international scope of the evidence base. Although the review was originally conceived with a particular interest in vocational education in the UK, the limited body of UK-based research

led to the inclusion of studies from a range of international contexts and time periods. However, vocational education systems may differ substantially across countries. These differences may limit the direct comparability of findings and their transferability beyond the specific settings in which individual studies were conducted. As a result, drawing consistent conclusions across this body of literature is challenging, and while the review aimed to draw out implications for vocational education in the UK, the findings should be interpreted with caution.

In addition, a substantial portion of the international literature is relatively dated, particularly the reviews of US studies from the 1930-80s. Although these studies provide valuable insights into broad patterns of predictive validity in relation to success in vocational education, the nature of vocational programmes and assessment practices has evolved considerably over time. Findings from older studies should therefore be interpreted with care and used primarily to identify general trends rather than to inform current policy.

There was also considerable variability in how predictors were assessed, along with differences in how success was measured (e.g., grades, course completion), which may limit the generalisability of results. Furthermore, certain variables used as predictors may have changed over time and thus may have not been conceptualised or measured consistently. For example, definitions of intelligence have evolved over time, with modern psychological conceptions shifting away from valuing practical, contextualised, and socially integrated skills towards greater emphasis on abstract, decontextualised cognition, speed, and novelty (Greenfield, 2020). This suggests that intelligence measures used in early studies may not fully reflect and capture today's definitions and constructs of intelligence. These variations highlight the need for more standardised research specifically focused on vocational qualifications, education, and training.

Some of the studies included in this review primarily examined predictors of success in academic qualifications (e.g., GCSEs) and were extrapolated to vocationally relevant subjects (e.g., Art and Design, Music). This approach provided insights into domains that overlap across qualification types, but it does not fully capture the unique structures, assessment formats, and purposes of vocational qualifications.

A final limitation concerns the search process. Although the scoping review followed a pre-specified protocol, it remains possible that some relevant studies were not identified. This may be due to the set of search terms used, and perhaps alternative methodology may have yielded additional relevant studies.

Discussion and conclusion

This scoping review set out to synthesise evidence on predictors of success in vocational qualifications, education, and training, and highlights some clear predictor-level patterns. Across the literature reviewed, prior academic attainment and general cognitive abilities remain the most consistent predictors overall, however, their predictive strength seems to vary substantially across vocational domains. Some non-cognitive traits, particularly Conscientiousness, also emerged as robust predictors of attainment, persistence, and satisfaction in

a range of vocational programmes, and in some cases, outperformed cognitive ability measures. These findings suggest that predictive validity in vocational education may depend on a nuanced, multi-layered set of predictors rather than any single indicator.

Across studies, prior academic attainment (e.g., GCSE performance or GPA) tends to emerge as the strongest single predictor across qualifications, although its effect is generally weaker for vocational than for academic routes (Barnard & Dixon, 1998). Correlations between prior attainment and later success also differ by subject. They tend to be stronger for more technical or business-oriented programmes and weaker for creative or performance-based domains such as Art and Design (Deary et al., 2007).

This pattern is mirrored in studies of cognitive abilities. General intelligence tends to be reported as a strong predictor of future achievement, although its strength as a predictor seems to be subject-dependent. Specific cognitive abilities – verbal, numerical, and non-verbal reasoning – have been identified as strong predictors of future success, but this is also based on how well they align with the vocational subject. Numerical ability tended to be more strongly related to technical and engineering courses (Bertua et al., 2005; Stock & Pratzner, 1969), spatial ability served as a good predictor in design-oriented and drafting subjects (Barnard & Dixon, 1998), and verbal ability produced the strongest correlations with business and clerical programmes (Prediger et al., 1968; Stock & Pratzner, 1969). In contrast, practical and creative subjects often showed weaker links with standard academic and cognitive measures (Deary et al., 2007), while non-verbal abilities tended to show the lowest predictive value overall. These findings emphasise that the strength of intelligence test scores as predictors of future success varies by subject, and that the predictive power of a single cognitive variable is not generalisable across all vocational areas. Perhaps using measures that reflect the skills, modes of learning, and assessment formats required in specific vocational areas as predictors would likely be more informative than relying solely on generic academic or intelligence indicators.

Prediger et al. showed how IQ, operationalised as a verbal intelligence measure (V-INTEL), was only one of the two best predictors of success for three of the 11 courses they reviewed, leading them to conclude that utilising IQ as the single predictor of success in vocational courses would "be very unfair to many students" (1968, p. 143). The varying correlations of predictors based on the different vocational subjects serve as further evidence that predictive validity hinges on the match between the ability construct being measured and the skills and assessment demands of the specific vocational subject. Domain-specific abilities that align with course content in some cases can predict success as well as, or better than, more general measures of prior attainment or intelligence.

Non-cognitive attributes have also emerged as significant predictors of future achievement in vocational education. With regard to personality traits, Conscientiousness (one of the Big Five personality traits) emerged consistently as a strong and robust predictor of attainment, persistence, and satisfaction across a range of vocational programmes, in some cases serving as a stronger predictor of success than prior attainment or cognitive ability. It is interesting to note that

in a meta-analysis on general education (not specific to vocational education), Conscientiousness has also been identified as the most consistent non-cognitive predictor of achievement (Poropat, 2009), with smaller, more context-sensitive effects for other traits.

Extraversion occasionally emerged as a predictor in People-focused occupations (Nießen et al., 2020), whereas the other personality traits (Openness, Agreeableness, and Emotional Stability) showed much weaker and less consistent associations. The use of vocational interests seemed to also play some role in predicting success and satisfaction. While generally they were reported as weaker predictors, vocational interests contributed to stronger and more reliable predictions of grades, satisfaction, and early apprenticeship success when combined with other predictors, such as cognitive ability. Motivation, engagement, and social-emotional competencies also featured in a number of studies. These variables were found to provide small, yet significant, contributions to predicting future success in vocational education. Creative potential, practical-scientific competencies, and manual dexterity have so far shown only weak and highly context-dependent associations with vocational outcomes, suggesting that they are less central as predictors compared to prior attainment, cognitive abilities, Conscientiousness, vocational interests, and other non-cognitive factors.

Beyond predicting achievement outcomes, several predictors, including prior attainment, reading ability, personality, and cognitive ability, were identified as also linked to students' likelihood of selecting an academic or vocational pathway. As a result, the profile of students on vocational pathways is likely to be different from that of students on academic pathways in terms of prior attainment and non-cognitive traits.

From the findings of this review, it seems that there is no single variable that can consistently predict success across vocational domains. Instead, vocational achievement seems to be shaped by combinations of prior academic achievement, cognitive abilities, and non-cognitive traits, with the relative contribution of each predictor varying by subject domain and programme type. The predictive value of these variables is strengthened when they align closely with the specific requirements of a vocational subject.

This suggests that measures such as prior attainment or cognitive tests may provide useful information for understanding whether students are likely to be well suited to meet the demands of particular vocational courses, depending on the nature of those demands. However, such measures do not capture the full range of abilities required for success in vocational education and there may be benefits to considering wider evidence about learners when making course entry decisions.

The findings of this review support a multidimensional view of vocational success, in which combinations of different types of predictors – such as prior attainment, cognitive abilities, and non-cognitive traits (including personality traits, vocational interests, and motivation) – are considered when examining students' likelihood of success in vocational education.

Future research could focus explicitly on current vocational qualifications and programmes, examining predictors of success within vocational pathways in their own right, rather than aggregating vocational and academic routes within a single analysis. Such work would help clarify which factors are most relevant for success in vocational education, while acknowledging that distinctions between vocational and academic pathways are not always clear-cut. Multivariate designs that include prior attainment, multiple cognitive abilities, personality traits, vocational interests, and motivational variables would help clarify their unique and combined contributions to different vocational outcomes (e.g., grades, completion, progression). Given the domain-specific patterns observed in this review, further work would also benefit from more fine-grained analyses by vocational subject or area. In particular, the development or evaluation of predictors that are better aligned with the skills and assessment demands of particular programmes may be especially informative. Finally, it may be useful to further investigate relatively less studied constructs such as creative thinking and behavioural engagement as potential predictors of success in vocational education.

In conclusion, predicting success in vocational education may be somewhat more challenging than predicting success in academic qualifications, but findings here suggest there may still be good opportunities to do so. Rather than relying on a single indicator, drawing on a broad set of predictors, and attending to how these relate to specific vocational pathways, has the potential to provide more informed guidance, target support, and improve understanding of progression for young people taking a vocational pathway in education.

References

Baird, J., James Relly, S., & Görgen, K. (2024). El caso de Inglaterra: Un problema de reconocimiento en la evaluación técnico-profesional [The case of England: Vocational assessment's recognition problem]. In M. T. Florez Petour, T. R. Rozas Assael, J. M. O. Astorga, & F. Hevia (Eds.), *Evaluación educativa y justicia social. Desafíos y horizontes desde una perspectiva internacional* (Capítulo 4). Casa Chata.

Barnard, P., & Dixon, S. (1998). Value added: Beyond A-levels to vocational programmes? *FEDA Bulletin, 2*(4), n4.

Baron-Boldt, J., Schuler, H., & Funke, U. (1988). Predictive validity of school grades: A meta-analysis. *Zeitschrift für Pädagogische Psychologie, 2*(2), 79–90.

Bertua, C., Anderson, N., & Salgado, J. F. (2005). The predictive validity of cognitive ability tests: A UK meta-analysis. *Journal of Occupational and Organizational Psychology, 78*(3), 387–409.

Brandt, N. D., Lechner, C. M., Tetzner, J., & Rammstedt, B. (2020). Personality, cognitive ability, and academic performance: Differential associations across school subjects and school tracks. *Journal of Personality, 88*(2), 249–265.

Chen, P., Goncharova, A., Pilz, M., Frommberger, D., Li, J., Romanova, O., & Lin, Y. (2021). International curriculum comparison in Vocational Education and Training: A collaborative development of an analysis instrument. *International Journal of Research in Vocational Education and Training, 8*(4), 16–43.

Conlon, G. (2005). The determinants of undertaking academic and vocational qualifications in the United Kingdom. *Education Economics, 13*(3), 299–313.

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35*(1), 13–21.

Department for Education. (n.d.). *Vocational technical qualifications (VTQs)*. Skills for Careers.

Diedrich, J., Neubauer, A. C., & Ortner, A. (2018). The prediction of professional success in apprenticeship: The role of cognitive and non-cognitive abilities, of interests and personality. *International Journal for Research in Vocational Education and Training, 5*(2), 82–111.

Dubeau, A., Plante, I., & Frenay, M. (2017). Achievement profiles of students in high school vocational training programs. *Vocations and Learning, 10*(1), 101–120.

Dubeau, A., Plante, I., Jutras-Dupont, C., Samson, G., & Frenay, M. (2021). Understanding the relationships between psychological and contextual determinants, motivation and achievement outcomes for students in vocational training or technical training programs. *Vocations and Learning, 14*(1), 165–183.

Erikson, R., & Jonsson, J. O. (1996). Explaining class inequality in education: The Swedish test case. In R. Erikson, & J. O. Jonsson (Eds.), *Can education be equalized?* (pp. 1–63). Westview Press.

Esser, H. (2016). The model of ability tracking – Theoretical expectations and empirical findings on how educational systems impact on educational success and inequality. In H. P. Blossfeld, S. Buchholz, J. Skopek, & M. Triventi (Eds.), *Models of secondary education and social inequality* (pp. 95–114). Edward Elgar Publishing.

Farnsworth, P. R. (1935). Are "music capacity" tests more important than "intelligence" tests in the prediction of the several types of music grades? *Journal of Applied Psychology, 19*(3), 347–350.

Ghiselli, E. E. (1966). *The validity of occupational aptitude tests.* John Wiley & Sons.

Gilbert, W., Stack, D. M., Barker, E. T., Dubeau, A., Serbin, L. A., & Véronneau, M. H. (2024). Emotional competence and help-seeking intentions as predictors of educational success in vocational training students. *Journal of Adult and Continuing Education, 30*(2), 617–639.

Greenfield, P. M. (2020). Historical evolution of intelligence. In R. J. Sternberg (Ed.), *The Cambridge Handbook of Intelligence* (pp. 916–939). Cambridge University Press.

Jellis, C. (2024). A short history of the Centre for Evaluation and Monitoring (CEM). *Research Matters: A Cambridge University Press & Assessment publication, 37*, 74–83.

Jencks, C., Smith, M., Acland, H., Bane, M. J., Cohen, D., Gintis, H., Heyns, B., & Michelson, S. (1979). *Inequality: A reassessment of the effect of family and schooling in America*. Basic Books.

Klapproth, F., & Schaltz, P. (2014). The validity of predictors of academic and vocational-training achievement. In S. Krolak-Schwerdt, S. Glock, & M. Böhmer (Eds.), *Teachers' professional development: Assessment, training and, learning* (pp. 153–166). Sense Publishers.

Kwalwasser, J. (1936). The composition of musical ability. *Teachers College Record: The Voice of Scholarship in Education, 37*(10), 35–42.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*(2), 175–215.

McGrath, S. (2012) Vocational education and training for development: A policy in need of a theory? *International Journal of Educational Development, 32*(5), 623–631.

Mitton, L., & Hensby, A. (2024). Bringing vocational qualifications into the inclusivity agenda: The case of the BTEC. In A. Hensby, & B. Adewumi (Eds.), *Race, capital, and equity in Higher Education: Challenging differential academic attainment in UK universities* (pp. 115–140). Palgrave Macmillan.

Molnár, G., & Kocsis, Á. (2023). Cognitive and non-cognitive predictors of academic success in higher education: A large-scale longitudinal study. *Studies in Higher Education, 49*(9), 1610–1624.

Nießen, D., Danner, D., Spengler, M., & Lechner, C. M. (2020). Big five personality traits predict successful transitions from school to vocational education and training: A large-scale study. *Frontiers in Psychology, 11*, 1827.

O'Rourke, L. J. (1938). Vocational aptitude tests. *Review of Educational Research, 8*(3), 257–268.

Ofqual. (2021, December 16). *Delivery and award of vocational and technical qualifications in 2021*. Ref: Ofqual/21/6874.

Ofqual. (2024). *Equalities analysis interactive report*. Ofqual.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S.,...Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ,* 372, n71.

Peters, M. D. J., Godfrey, C. M., Khalil, H., McInerney, P., Parker, D., & Soares, C. B. (2015). Guidance for conducting systematic scoping reviews. *International Journal of Evidence-Based Healthcare, 13*(3), 141–146.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*(2), 322–338.

Prediger, D. J., Waple, C. C., & Nusbaum, G. R. (1968). Predictors of success in high school level vocational education programs: A review, 1954-1967. *The Personnel and Guidance Journal, 47*(2), 137–145.

Samuelson, C. O. (1958). Interest scores in predicting success of trade school students. *The Personnel and Guidance Journal, 36*, 538–541.

Schofield, C., & Dismore, H. (2010). Predictors of retention and achievement of higher education students within a further education context. *Journal of Further and Higher Education, 34*(2), 207–221.

Stock, W. E., & Pratzner, F. C. (1969). *Review of research on student selection and the prediction of success in occupational education.* University of Minnesota.

Tricco, A. C., Lillie, E., Zarin, W., et al. (2016). A scoping review on the conduct and reporting of scoping reviews. *BMC Medical Research Methodology, 16*(15).

UCAS. (n.d.). *UCAS Tariff points.*

Usslepp, N., Hübner, N., Stoll, G., Spengler, M., Trautwein, U., & Nagengast, B. (2020). RIASEC interests and the Big Five personality traits matter for life success – But do they already matter for educational track choices? *Journal of Personality, 88*(5), 1007–1024.

Vasalampi, K., Tolvanen, A., Torppa, M., Poikkeus, A. M., Hankimaa, H., & Aunola, K. (2023). PISA reading achievement, literacy motivation, and school burnout predicting adolescents' educational track and educational attainment. *Learning and Individual Differences, 108*, 102377.

VERBI Software. (2025). *MAXQDA 24.7.0* [Computer software].

Vidal Rodeiro, C., & Vitello, S. (2020). *Vocational qualifications at Key Stage 4 and Key Stage 5: Who takes them and how they fit into students' programmes of study.* Cambridge Assessment.

Vitello, S., & Williamson, J. (2017). Internal versus external assessment in vocational qualifications: A commentary on the government's reforms in England. *London Review of Education, 15*(3), 536–548.

Volodina, A., Nagy, G., & Köller, O. (2015). Success in the first phase of the vocational career: The role of cognitive and scholastic abilities, personality factors, and vocational interests. *Journal of Vocational Behavior, 91,* 11–22.

Williams, M. E., Clarkson, S., Hastings, R. P., Watkins, R. C., McTague, P., & Hughes, J. C. (2022). Factors from middle childhood that predict academic attainment at 15-17 years in the UK: A systematic review. *Frontiers in Education. 7*, 849765.

Williamson, J. & Carroll, M. (2018, November 7–10). *Vocational, technical and applied qualifications within upper secondary education in England: Who takes them, how do they fit within students' programmes of study, and where do students progress to next?* [Conference presentation]. 19th annual AEA-Europe conference, Arnhem/Nijmegen, The Netherlands.

Winch, C. (2021). Learning outcomes: The long goodbye: Vocational qualifications in the 21st century. *European Education Research Journal, 22*(1), 20–38.

Wyness, G., Macmillan, L., Anders, J., & Dilnot, C. (2023). Grade expectations: How well can past performance predict future grades? *Education Economics, 31*(4), 397–418.

Appendix: Terms used in literature searches

The following search terms and combinations were used during the literature search:

- Predictive validity AND vocational qualification
- Predictive validity AND grade
- Predictors AND vocational qualifications
- Vocational predictive validity AND VQs
- Correlation predictive validity vocational qualification
- Regression predictive validity vocational qualification
- Meta analysis AND predictive validity AND vocational qualification
- Predictors of vocational qualification success
- Correlation predictors of vocational qualifications
- Predictors of vocational qualification
- Vocational qualification AND predictive validity
- Correlation predictors AND VQ success
- Predictive validity reasoning tests
- Vocational success predictors
- Predictors AND vocational

How does examining influence teachers' assessment literacy?

Victoria Coleman (Research Division)[1] **and Martin Johnson** (Research Division)

Introduction

Assessment literacy (AL), which includes teachers' conceptualisation of assessment, is an important aspect of teacher professionalism. AL influences teachers' assessment practices (Deneen & Brown, 2016), and affects student outcomes (Mellati & Khademi, 2018). Although it is increasingly highlighted as a key component of teacher professional competency, there are concerns that many teachers have only limited AL and that this can influence their understanding of examinations, and their allied teaching practices (Department for Education, 2015; National Association of Head Teachers, 2014). The examination system in the UK largely relies on teachers choosing to work as examiners who mark examination scripts for external awarding bodies (Menzies, 2015; Royal-Dawson & Baird, 2009). This practice of using teachers to mark qualifications is also common in other countries (Royal-Dawson & Baird, 2009). With teachers choosing to examine on top of an already busy teaching schedule (Menzies, 2015), one motivating factor may be that examining helps teachers to develop their AL.

What is assessment literacy (AL)?

AL is a complex and expansive concept and has been prone to change over time. As a preliminary step to exploring teacher-examiners' AL, we felt it was important to capture contemporary understandings of the concept as reflected in the academic literature. This then allowed us to capture our own definition of AL and to use this as a basis for gathering teachers' reflections in our research project.

> AL refers to a teacher's ability to use assessment effectively as part of their practice. It includes the knowledge and skills that a teacher uses to gather and interpret information about student achievement and encompasses a teacher's beliefs and feelings about the value and purpose of assessment. AL is influenced and mediated by external factors such as the teacher's experiences of assessment and the environments in which they develop and apply their AL.

Figure 1: Our definition of AL

1 Victoria led this research while working in the Research Division at Cambridge University Press & Assessment.

Our definition of AL (Figure 1) refers to teachers' abilities to use assessment effectively as part of their practice and includes their knowledge and skills around gathering and interpreting information about student achievement. This aspect of the definition reflects some traditional conceptualisations of AL which have tended to focus primarily on the knowledge and skills teachers need to apply summative assessment measures (Coombs, 2017) but which have now evolved to subsume formative and student-centred assessment approaches (DeLuca et al., 2016).

Our definition of AL also recognises that it is affective and dynamic, involving teachers' beliefs and feelings about the value and purpose of assessment. Building on the idea that AL is changeable over time (Looney et al., 2017; Xu & Brown, 2016), our definition recognises that AL is influenced by a person's environment and experience. In this way it is possible to conceive that teachers have an "assessment career" (Deneen & Brown, 2016), which starts when they are themselves students (through being subjected to assessment by others), develops as they engage in teacher training, and continues as they apply assessment during their professional work. This aspect of AL highlights how it is influenced by factors that are external to a teacher (Fulmer et al., 2015), with some of these factors occurring at the micro level (e.g., classroom context), the meso level (e.g., school policies), and the macro level (e.g., national education policies).

Teacher AL in the UK

Concerns have been raised that many teachers do not have sufficient levels of AL, which is problematic given that teacher AL is an important component of a successful assessment system (National Association of Head Teachers, 2014), and impacts student learning and achievement (Popham, 2009; Stiggins, 2010). Estimations suggest that a typical teacher spends up to a third of their professional time on assessment-related activities (Stiggins, 2010). In the UK, many teachers lack confidence in their assessment ability, with only 35 per cent of classroom teachers describing themselves as feeling very confident conducting day-to-day assessment as part of their teaching (Millard et al., 2017). In addition, assessment practice and governance have been identified as areas requiring improvement across UK schools (Ofsted, 2022).

The UK Department for Education (DfE) has detailed eight Teacher Standards which specify the minimum level of practice expected of trainee teachers, newly qualified teachers, and teachers working in maintained schools. As part of this, *Standard 6* specifies the assessment requirements for teachers (Department for Education, 2021), with Initial Teacher Training (ITT) providers needing to guarantee that students leaving their programmes meet these Standards. These arrangements arose partly as a response to concerns that assessment was not taught to a sufficient standard in UK ITT courses (e.g., see Department for Education, 2015; Millard et al., 2017; National Association of Head Teachers, 2014). However, some commentators have continued to express concerns that this initial teacher training model requires reinforcement from "a three-or-more-year structured package of support, including mentoring from 'expert colleagues' … with appropriate knowledge and understanding of assessment theory and practice – attributes that are currently in short supply" (Walker, 2021).

Teacher-examiners and AL

The examining system in the UK largely relies on teachers working as examiners to mark scripts, and this practice is also common in other countries and states, including Canada, France, New Zealand, Australia, and the US (Royal-Dawson & Baird, 2009). In the UK, teachers can apply to become examiners. As part of this role they receive training and support prior to and during examining (Joint Council for Qualifications, 2026), which includes their participation in standardisation activities that help to ensure that they can apply mark schemes consistently. Although it is possible that these experiences act as a form of assessment Continuing Professional Development (CPD), consideration of whether examining is beneficial to teachers' professional development is underexplored by research.

Research by both Menzies (2015) and Lockyer (2018) found that teachers mainly chose to examine to improve their teaching practice and to benefit their students. Moreover, this benefit often related to assessment needs, for example, Lockyer reported that the most commonly given reason for beginning examining was "to better prepare my students for assessment". In addition, both studies found that examiners did not feel they were paid sufficiently, suggesting that teachers choose to examine for reasons beyond monetary gain.

Taken together, these findings may be interpreted to suggest that some teachers' choices to examine primarily related to improving their understanding of assessment. In this sense, it could be the case that examining practice is performing an assessment CPD function.

While there has only been limited research on the impact of examining on teacher AL, some studies highlight the importance of professional collaboration and training as mechanisms for developing teacher AL. Johnson and Shaw (2012) found that teachers with examining experience felt that they had insights into the examination system that non-examining teachers did not have. The authors concluded that the teachers' participation in an examiner community of practice gave them the cognitive support to enable them to make sense of the language and codes used by other examiners. In Australia, research into the impact of participating in online moderation meetings on teachers found that the meetings provided an opportunity for professional conversations about assessment and that this helped teachers to develop their assessment identity (Adie, 2013).

Research focus

This study built on earlier research findings (Johnson & Coleman, 2021) and aimed to explore whether teachers who were also examiners had a common understanding of AL, and what their conceptualisation of AL included. We also wanted to investigate teachers' perceptions of the impact that examining had on their AL, how this may have influenced their professional practice and relationships, and which aspects of the examining experience influenced the development of AL.

Methods and materials

While there are a number of existing tools for evaluating teacher AL, these have a number of disadvantages, such as having a limited perspective on what actually counts as AL (DeLuca et al., 2016), or constraints around eliciting tacit AL from teachers (Sadler & Reimann, 2018). To explore our research questions and to overcome limitations in the availability of measurement instruments or guiding theory (Creswell & Clark, 2011) we used a two-phase exploratory mixed-methods design. Both phases of this research followed our institution's research ethics procedures and guidelines, including obtaining informed consent and ensuring confidentiality. This article focuses on the outcomes of the second of these study phases.

Our first exploratory research project phase (reported in Johnson & Coleman, 2021) used a narrative concept mapping approach with a group of eight teacher-examiners who had worked with two awarding organisations. Narrative concept mapping methods centre on the discussions that emerge from participants' reflections on their generation of concept maps. Our narrative concept mapping process involved the teacher-examiners carrying out three tasks. Task 1 was a 10-minute warm-up brainstorm exercise asking them to respond to the question: "What words come into your mind when you think about the word 'Assessment'?" For Task 2 we then introduced the teacher-examiners to our definition of AL (Figure 1) and to the concept map idea. The teacher-examiners then completed a concept map around the question of "How has examining developed or changed your assessment literacy?" After completion of the concept map, Task 3 involved a semi-structured interview that included questions around (i) their role as an examiner, (ii) their definition of AL, and (iii) questions around their concept maps to explore the contribution that examining had on their AL and wider practice. For more information on the method used in this phase of the project see Johnson and Coleman (2021).

We then used the outcomes of this exploratory phase as the basis for a second, confirmatory phase where we developed a survey to explore AL across a larger sample of teacher-examiners. Our analysis of this second phase survey data is the focus of the current research article.

We used an opportunity sampling approach, contacting members of the teacher-examiner community to see if they would be interested in participating in our research. We distributed our survey as a web link through personal and organisational social networks, and it was available for eight weeks. We offered a prize draw voucher as encouragement for participation. The survey was designed to capture mainly objective, closed-choice responses which could be analysed quantitatively, although most included an option for the participant to input additional information, if needed.

We gathered information about each participant's professional history, and we screened them to ensure that they were currently teaching and that they had recent examining experience (i.e., within the last five years). After probing their notion of AL, the teacher-examiners were presented with our definition of AL and asked to consider the nature and magnitude of the influences that they perceived

to have affected their own AL (with specific probes focusing on the influence of examining).

There were 103 survey respondents, and these were teacher-examiners primarily based in the UK. Not all teacher-examiners completed every question in the survey, with some skipping questions or stopping part-way through. Of those who gave information about their gender, 61 per cent were female and 39 per cent were male.

The teacher-examiners had been teaching for between 2 and 41 years (mean = 17.4 years, standard deviation = 8.95), and their examining experience ranged from 6 months to 30 years (mean = 6.4 years, standard deviation = 5.86). The participants were asked to list the subjects that they were teaching at the time of the survey. We then grouped the subjects into overarching areas, e.g., subjects such as Sport Science and Health Education were grouped into Science, and subjects such as Psychology and Law were grouped into Social Science (see Appendix for full list).

Of those surveyed, 66 per cent of the participants were teaching just one subject, while the remainder listed two or more subjects. The teacher-examiners were also asked about the qualifications they were examining, with 40 per cent of the participants examining one qualification, 34 per cent examining two qualifications, and 25 per cent examining three qualifications. English subjects were the most commonly reported examined subject (28 per cent), followed by Science subjects (24 per cent), and Social Science subjects (20 per cent) (see Figure 2 for full information on the subjects taught and examined).

The teacher-examiners were asked about the levels of education that they were teaching at, which were phrased in terms of Key Stages (KS) used in the English system (Roberts, 2019). Of the responding teacher-examiners, 23 per cent taught at only one level, with the rest teaching at two or more levels. The most commonly taught levels were KS5 (80 per cent) and KS4 (75 per cent), which align with key national examination points in the UK. Most of the qualifications being examined were GCSE / IGCSE[2] (42 per cent) and GCE A Level[3] / IB diploma[4] qualifications (35 per cent), although there were a variety of other qualifications reported such as vocational or technical qualifications, and English for Speakers of other Languages (ESOL) qualifications.

2 General Certificate of Secondary Education (GCSE) is an examination usually taken at the end of compulsory schooling (KS4) in England, Wales, and Northern Ireland. International General Certificate of Secondary Education (IGCSE) is an international high school qualification that prepares students for further study.

3 General Certificate of Education (GCE) A Level is usually studied over a two-year period and is widely recognised in England, Wales and Northern Ireland as being the standard entry qualification for assessing the suitability of applicants for academic courses in UK universities.

4 International Baccalaureate (IB) diploma is a two-year academic programme. It is at the same level as GCE A Level qualifications and can also be used for entry for academic courses in UK universities.

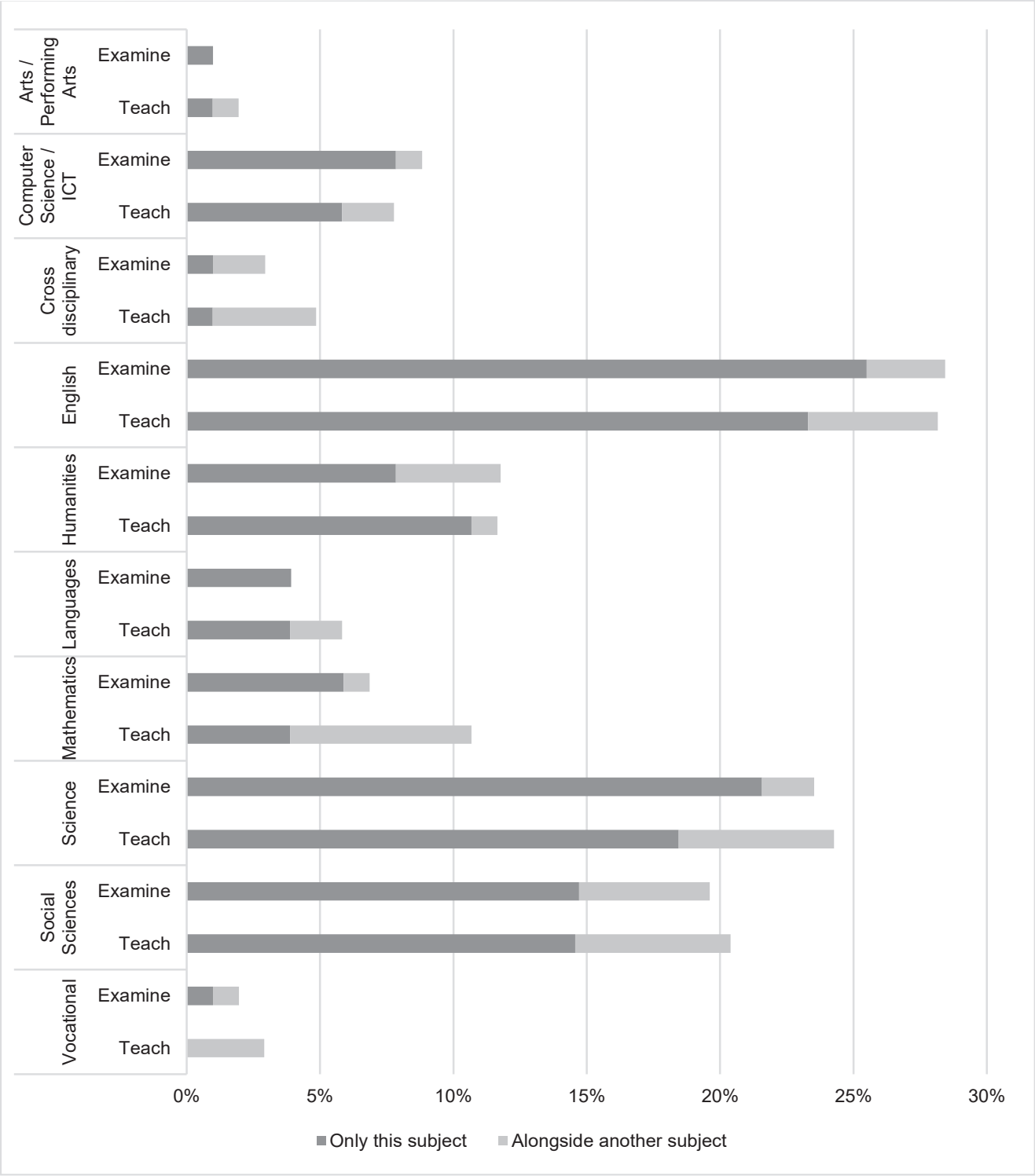

Figure 2: Subjects taught and examined[5]

Results

Understandings of AL

Only 28 per cent of the teacher-examiners were familiar with the term "assessment literacy". Most of these participants considered AL specifically in relation to summative assessment, for example, they associated it with being able

5 ICT refers to "Information and Communication Technology".

to interpret summative assessment results, knowing test instruction conventions, knowing the standards for test development, and being aware of the impact of assessments on test takers.

Teacher-examiners were then presented with our definition of AL and asked, "On a scale of 1 to 5, how assessment literate do you feel you are?" (with 1 being "not at all" and 5 being "extremely"). The ratings of the 100 teacher-examiners who responded to this question showed that they felt that they had well-developed AL (mean = 4.3, standard deviation = 0.62). No respondents gave themselves a rating below 3.

Influence of experience on AL

Having been presented with our definition of AL, the teacher-examiners were asked to reflect on the relative influence of four different experiences (being examined, Initial Teacher Training, teaching, and examining) on their AL using a 1 to 5 scale (with 1 being "has not influenced" and 5 being "has greatly influenced"). This question was answered by 102 respondents, and examining was rated as the most influential experience (Table 1). All teacher-examiners reported that examining had influenced their AL, with the majority (77 per cent) selecting a rating of 5. Teacher-examiners were also able to give and rate additional experiences that they considered as having influenced their AL. A number of teacher-examiners indicated that their AL was influenced by other non-examining related CPD, other assessment roles (such as working on assessment design), other qualifications, and other career experiences. Taken together, these findings supported the notion that examining was very influential on AL, and that several diverse experiences could also influence an individual teacher-examiner's AL.

Table 1: Ratings of the impact of experiences on teacher-examiners' AL

Experiences	No. of participants	Mean	SD	Range
Examining	102	4.73	0.53	3-5
Teaching	102	3.49	0.96	1-5
Initial Teacher Training	102	2.51	1.31	1-5
Own experience of being examined	102	2.69	1.34	1-5
Other experiences	**No. of participants**	**Mean**	**SD**	**Range**
Other (non-specified)	14	3.36	0.93	2-5
Non-examining CPD	10	4.50	0.71	3-5
Other assessment role (e.g., test construction)	6	4.67	0.52	4-5
Participating in research on assessment	4	4.50	0.58	4-5
Other qualification	2	5.00	0.00	-
Other career experience	2	4.00	1.41	3-5
Training others in assessment	2	4.50	0.71	3-5
National performance standards	1	1.00	-	-
Seeing how other schools assess	1	1.00	-	-
Social media	1	1.00	-	-

Influence of examining on assessment knowledge and skills

In our earlier exploratory study, our teacher-examiner sample identified 24 aspects of AL knowledge and skills that they had gained through examining. Our current survey then asked the larger sample of teacher-examiners to indicate which knowledge and skills they felt had been influenced by their examining experience. This question was answered by 98 participants (Figure 3).

The teacher-examiners picked between 3 and 24 AL knowledge and skills aspects that were affected by their examining work (mean = 16.6, standard deviation = 5.72), suggesting that examining had influenced their AL quite widely. The aspect of AL knowledge and skills most influenced by examining was "understanding of mark schemes" (95 per cent), while the least was "ability to communicate assessment data" (44 per cent).

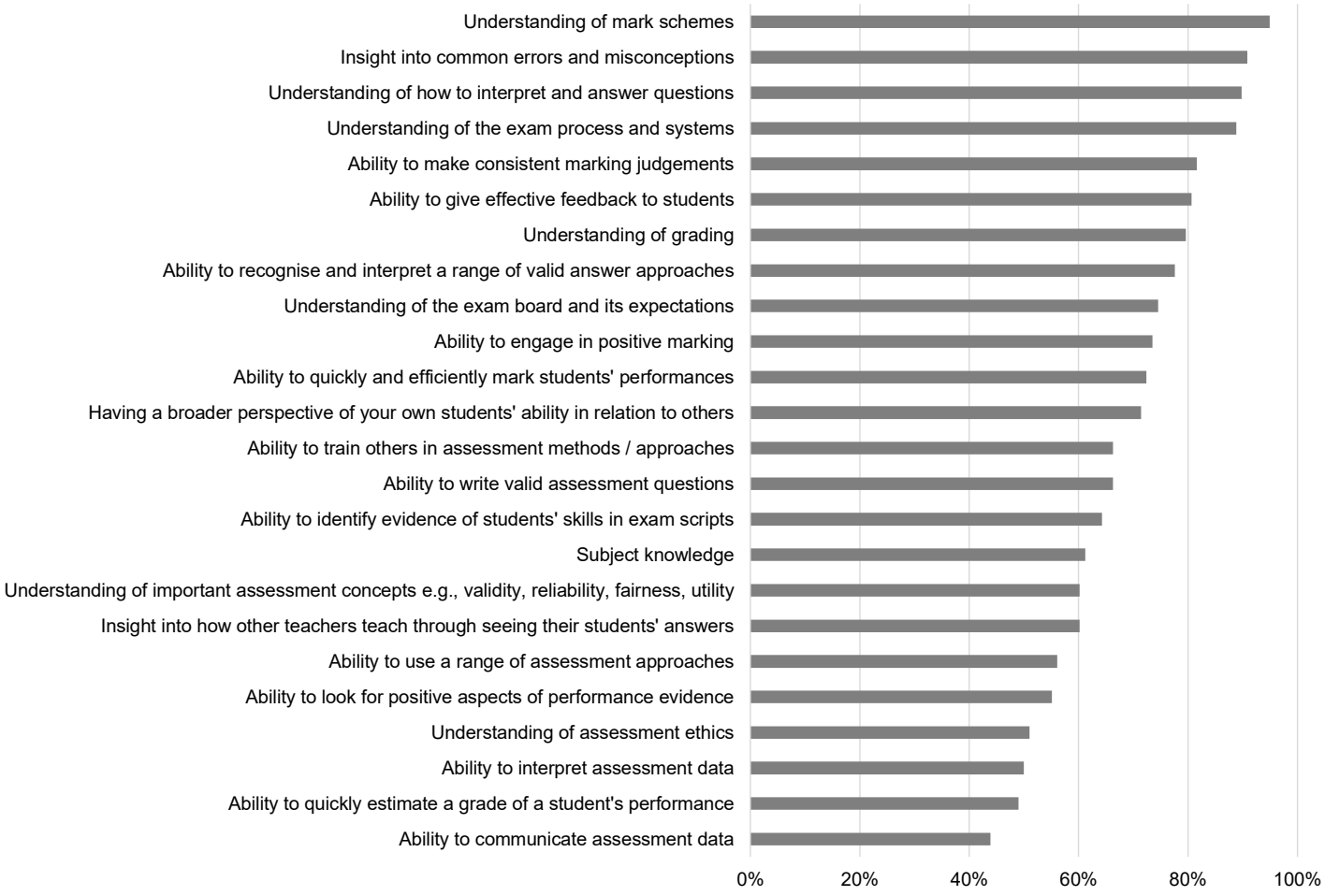

Figure 3: Aspects of AL influenced by examining

Influence of examining on assessment beliefs and feelings

In our earlier exploratory study, our teacher-examiner sample indicated that examining had influenced a variety of their beliefs and feelings. Our current survey then listed these and asked the larger sample of participants to indicate their current beliefs and feelings about assessment. These questions were answered by 94 teacher-examiners, who selected up to 14 options (mean = 6.6, standard deviation = 2.75). The participants were then asked whether examining had influenced the beliefs and feelings that they held, using a "yes / no" response. Figure 4 shows all the views held by teacher-examiners and the proportion who indicated that examining had influenced that view.

Most participants who reported holding a given belief or feeling felt that examining had influenced this view. The most common beliefs or feelings about assessment were positive, with the two most reported ("assessment is a measure of the application of knowledge", and "assessment is valuable") being held by approximately 75 per cent of teacher-examiners. Most respondents did not subscribe to some of the more negative beliefs or feelings about assessment (e.g., "assessment provides inaccurate information", and "assessment is bad for students and teachers").

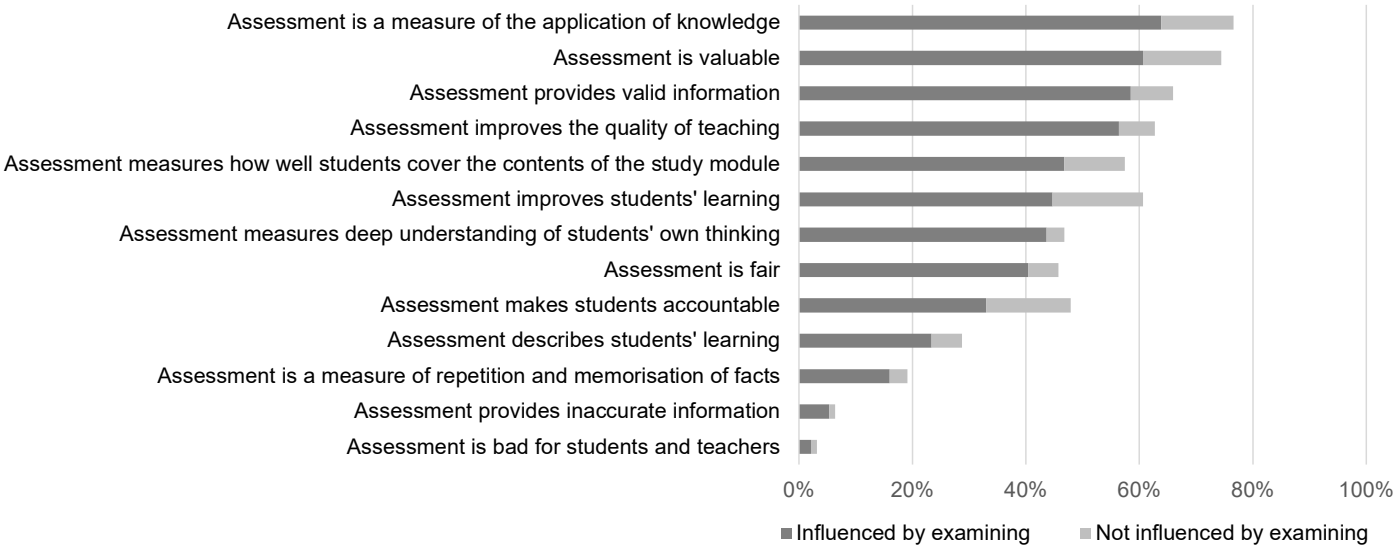

Figure 4: Beliefs and feelings about assessment

Influence of examining on teaching practice and professional relationships

We presented each teacher-examiner with the list of AL knowledge and skills that they had previously indicated had been affected by their examining. The participants then rated the extent to which those knowledge and skills aspects affected their classroom practice (using a scale of 1 to 5, with 1 being "not at all" and 5 being "greatly"). As shown in Figure 5, the most frequently selected rating for most aspects of AL was 5, with ratings of 1 or 2 being given infrequently.

The highest average rating was for "ability to give effective feedback to students" (mean = 4.6, standard deviation = 0.67), and the lowest was given for "insight into how other teachers teach through seeing their students' answers" (mean = 3.9, standard deviation = 1.06). Overall, where examining had influenced aspects of AL, it was considered to affect classroom practice to a great extent.

In the initial exploratory phase, themes relating to trust and the teacher-examiners' confidence in their AL emerged. To explore this further, the participants were asked in our current survey about the extent to which examining had influenced the trust they perceived others to have in them, and their self-confidence in their assessment ability. Of the 93 respondents who answered these questions, many teacher-examiners felt that examining had led to an increase in trust on matters of assessment from their students (80 per cent) and colleagues (70 per cent), but fewer (40 per cent) felt that it had influenced trust on matters of assessment from parents. In terms of the influence of examining on their self-confidence, a large proportion of the respondents felt that there had been an increase in their confidence in three areas: their ability to talk to students about assessment (95 per cent), their ability to apply mark schemes (92 per cent), and their knowledge of exam board expectations (88 per cent). The teacher-examiners were also asked whether there had been any influence of their examining at a school or departmental level. Of the 92 participants who responded to this question, 71 per cent agreed that their examining had an influence at a school or departmental level.

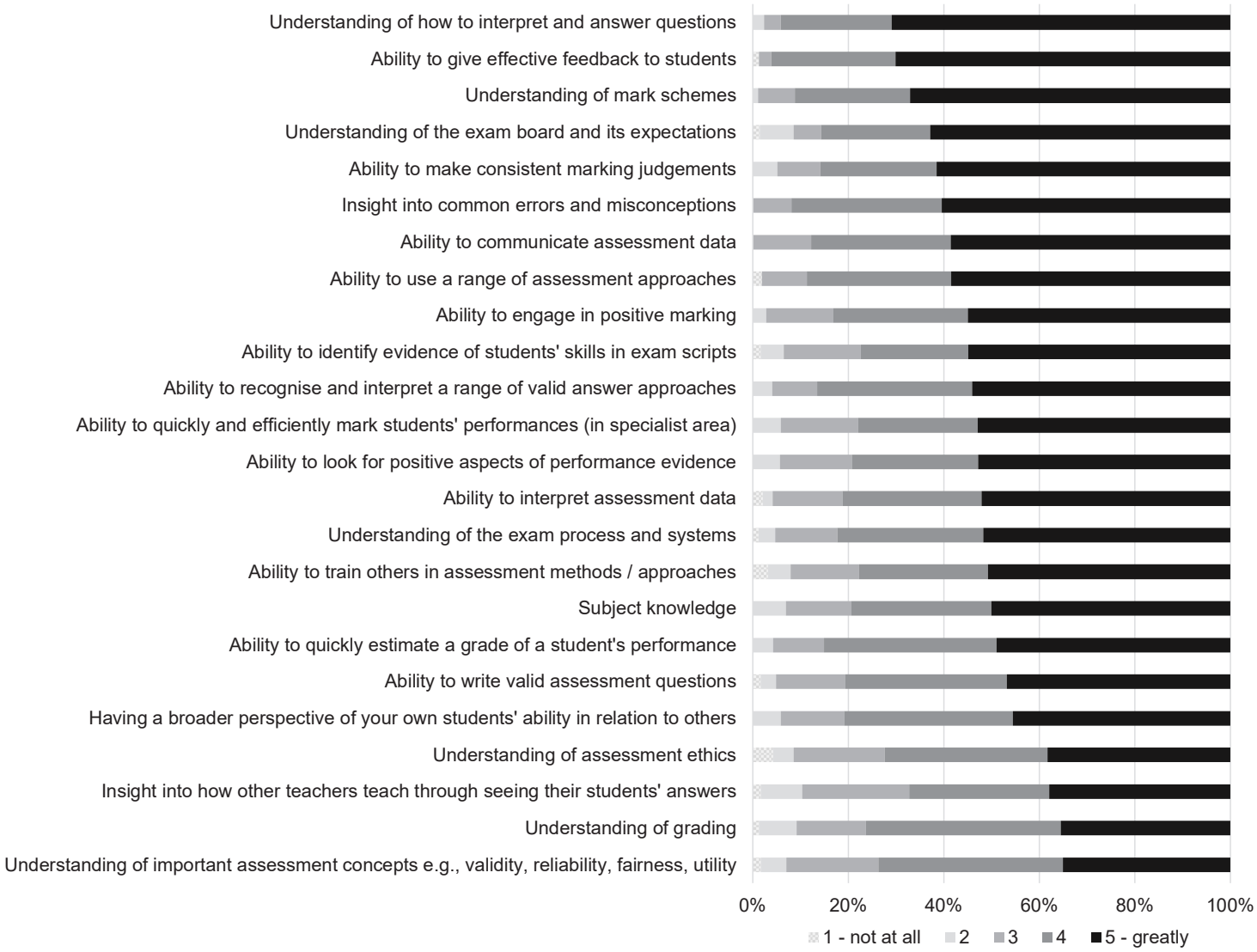

Figure 5: Influence of changes in AL on classroom practice

How does examining influence AL?

We were interested in how examining experience influenced the teacher-examiners' AL development. In our earlier exploratory study, participants identified 14 aspects of examining that influenced their AL, which could be grouped into three stages of the examining process: pre-marking preparation, marking, and post-marking activities. When we asked the participants in our current survey, 91 teacher-examiners selected a variety of aspects of examining that influenced their AL (mean = 8.5, standard deviation = 3.45), suggesting that these influences were linked to a variety of different experiences related to the examining process.

Figure 6 shows that marking-related activities – such as "applying mark schemes", "identifying common errors", and "marking large volumes of papers" – had the greatest influence on participants' AL. That said, all but two of the other aspects were reported as being influential by most of the participants. Moreover, although "contributing to the examiners' [post marking] report" was the least influential aspect of examining, it was still felt to have impacted AL for over 40 per cent of the teacher-examiners.

These results show that the influence of examining on AL is multifaceted, and that the benefits of examining for AL arise throughout the duration of the examining process, both related to and beyond the act of marking papers itself.

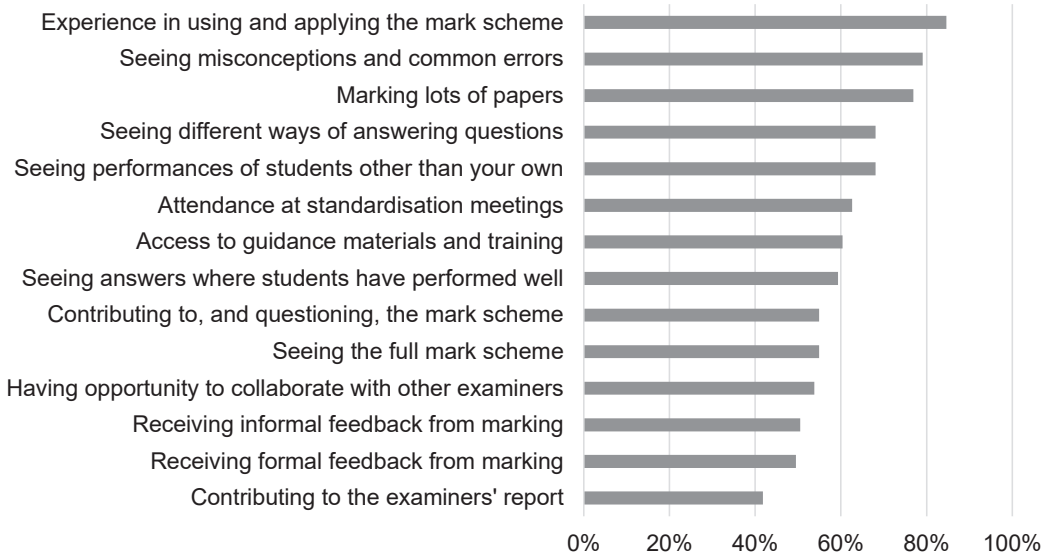

Figure 6: Aspects of examining that influenced AL

Discussion

In this research we examined teacher-examiners' understandings of AL and explored whether examining had impacted their AL, how this influenced their professional practice and relationships, and what aspects of the examining experience contributed to any change in their AL. Literature suggests that AL is a broad concept and is influenced by different experiences in a teacher's career. Moreover, while AL is considered to be very important for teaching practice, there is longstanding criticism that some ITT programmes do not equip teachers with sufficient AL. With the examination system in the UK reliant on teachers acting as examiners, we wanted to explore whether examining acted as a form of assessment CPD, providing teachers with an opportunity to develop their AL. Additionally we wanted to understand the breadth of this AL, and the impact it had on teachers' professional practice.

We found that most of the teacher-examiners were unfamiliar with the term AL. For those that were, definitions focused on summative assessment rather than the broad definitions which consider other forms of assessment. The lack of familiarity with the term AL suggests that it is an area where more work should be done to establish a common, professional, explicit understanding of AL, which goes beyond simply focusing on summative assessment.

When presented with our broad definition of AL (including knowledge, skills, beliefs, and feelings), most of the teacher-examiners considered themselves to be highly assessment literate and felt that their AL was significantly influenced by their examining work. While teaching was reported as a somewhat influential life experience on AL, the influence of ITT was the lowest rated of the listed life experiences. Although there might be recency effects acting on this finding (with the impact of contemporary teaching and examining being more accessible

experiences for reflection than historic training experiences), we need also to assume that the accrued experience of teacher-examiners would leave them well placed to consider the respective influence of different experiences in relation to each other. The teacher-examiners' reflections also seem to support the criticisms made by other observers that some ITT courses have not provided sufficient training on assessment (e.g., Department for Education, 2015; Millard et al., 2017; National Association of Head Teachers, 2014). This finding also supports the notion that examining can act as a form of CPD for teachers who may otherwise receive only limited assessment training.

The teacher-examiners reported that examining influenced their AL in terms of developing their knowledge and skills. Moreover, the teacher-examiners reported that this influence on their AL was broad. As well as influencing aspects that specifically related to summative assessment and examinations, they also reported that examining had influenced aspects with broader applicability to assessment and learning, such as more formative aspects of assessment related to giving feedback and spotting misconceptions. Similarly, many participants reported that examining had increased their AL in terms of foundational assessment concepts and skills. These findings therefore support the notion that examining experience exposes teachers to important assessment concepts that generalise across individual assessment episodes beyond summative examinations and give language to their assessment knowledge that can be used to connect with other teachers and assessment professionals.

Examining was reported to have influenced teacher-examiners' beliefs and feelings about assessment. These beliefs and feelings were multiple and diverse, showing that conceptions of assessment were complex and multifaceted. The teacher-examiners also generally reported having positive beliefs and feelings about assessment, with many of them indicating that examining had influenced these views. Although we anticipated that examining would influence teaching practice in terms of examination preparation, we were surprised at the extent to which the teacher-examiners reported that changes in their AL brought about by examining had influenced their wider classroom practice. This suggested that teacher-examiners believed that examining not only enabled them to be better at carrying out examining activities but also supported their subject teaching.

Teacher-examiners were asked to reflect on the impact of examining on their professional relationships. There was a clear impact of examining on perceived trust, with teacher-examiners reporting that they had benefitted from increased levels of trust from students and colleagues on matters of assessment. These findings highlight that the impact of examining on AL has a social dimension, with assessment expertise contributing to teachers' human capital – which itself has allied social status (Eide & Showalter, 2010). In terms of their own confidence in assessment matters, teacher-examiners reported that examining had been highly influential. While it is not possible to determine whether this enhanced confidence and trust were singularly related to the development of AL, it highlights that examining has an influence on teacher professional identity and on their relationships more broadly. Overall, the relationship between AL and teacher self-confidence and trust from others warrants future exploration.

Finally, most teacher-examiners reported that their examining had an influence at a school or departmental level. This has not yet been explored by research and suggests that examining as CPD can have benefits beyond improving an individual teacher's AL.

This study highlights that there are a variety of aspects of the experience of examining which influence teachers and their AL. The aspects of examining that were most commonly reported as having influenced AL related to the marking stage, particularly aspects related to seeing multiple exam responses. This suggests that it is the act of examining itself, rather than the training and guidance received, which had a significant impact on AL. This may be because examiner training and guidance is very focused on the specific examining work itself, while the act of marking and seeing lots of exam responses is an unintentional side effect of examining and has an impact upon AL more widely.

This study was exploratory using a small opportunity sample and some participants skipped questions. Consequently, our ability to generalise the findings to teacher-examiners more widely is limited. However, teacher-examiner AL has not yet been explored in detail, and so this research provides useful insight into how examining may have an impact on multiple aspects of teacher AL and classroom practice. Building on this, longitudinal research investigating teacher AL, and comparisons of examining and non-examining teachers, is needed to disentangle the impact of examining from other factors and understand how to harness examining to benefit teacher AL and teaching practice.

Overall, our findings highlight that AL is a complex and multifaceted concept, with teacher-examiners holding multiple knowledge types, skills, beliefs, and feelings about assessment. This study provides evidence that examining can play an important role in developing teacher AL, acting as a valuable form of assessment-related CPD. This suggests that more should be done to both understand and highlight the wider benefits of examining in terms of improving teacher AL and professional practice.

References

Adie, L. (2013). The development of teacher assessment identity through participation in online moderation. *Assessment in Education: Principles, Policy & Practice, 20*(1), 91–106.

Coombs, A. J. (2017). *Teacher educators' approaches to assessment* [Unpublished Master of Education thesis, Queen's University].

Creswell, J. W., & Clark, V. L. P. (2011). *Designing and conducting mixed methods research* (2nd ed.). Sage.

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016). Teacher assessment literacy: A review of international standards and measures. *Educational Assessment, Evaluation and Accountability, 28*(3), 251–272.

Deneen, C. C., & Brown, G. T. L. (2016). The impact of conceptions of assessment on assessment literacy in a teacher education program. *Cogent Education, 3*(1), 1225380.

Department for Education. (2015). *Carter Review of Initial Teacher Training*.

Department for Education. (2021). *Teachers' Standards: Guidance for school leaders, school staff and governing bodies*.

Eide, E. R., & Showalter, M. H. (2010). Human capital. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (3rd ed.), (pp. 282–287). Elsevier.

Fulmer, G. W., Lee, I. C. H., & Tan, K. H. K. (2015). Multi-level model of contextual factors and teachers' assessment practices: An integrative review of research. *Assessment in Education: Principles, Policy & Practice, 22*(4), 475–494.

Johnson, M., & Coleman, V. (2021). Out of their heads: Using concept maps to elicit teacher-examiners' assessment knowledge. *International Journal of Research & Method in Education, 44*(3), 257–272.

Johnson, M., & Shaw, S. (2012). Interpreting examiners' annotations on examination papers: A sociocultural analysis. *Irish Educational Studies, 31*(4), 467–485.

Joint Council for Qualifications. (2026). *Becoming an examiner*.

Lockyer, C. (2018). *Survey of examiners 2018: Headline findings*. (Ofqual/18/6449/6). Ofqual.

Looney, A., Cumming, J., van der Kleij, F., & Harris, K. (2017). Reconceptualising the role of teachers as assessors: Teacher assessment identity. *Assessment in Education: Principles, Policy and Practice, 25*(5), 442–467.

Mellati, M., & Khademi, M. (2018). Exploring teachers' assessment literacy: Impact on learners' writing achievements and implications for teacher development. *Australian Journal of Teacher Education, 43*(6), 1–18.

Menzies, L. (2015). *The future for examining*. LKMco/OCR.

Millard, W., Small, I., & Menzies, L. (2017). *Testing the water: How assessment can underpin, not undermine, great teaching.* LKMco/Pearson.

National Association of Head Teachers. (2014). Report of the NAHT Commission on Assessment. NAHT.

Ofsted. (2022). *The Annual Report of His Majesty's Chief Inspector of Education, Children's Services and Skills 2021/22*. Ofsted.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Into Practice, 48*(1), 4–11.

Roberts, N. (2019). *The school curriculum in England (Briefing Paper 06798)*. UK Parliament.

Royal-Dawson, L., & Baird, J.-A. (2009). Is teaching experience necessary for reliable scoring of extended English questions? *Educational Measurement: Issues and Practice, 28*(2), 2–8.

Sadler, I., & Reimann, N. (2018). Variation in the development of teachers' understandings of assessment and their assessment practices in higher education. *Higher Education Research & Development, 37*(1), 131–144.

Stiggins, R. (2010). Essential formative assessment competencies for teachers and school leaders. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of Formative Assessment* (pp. 233–250). Taylor & Francis.

Walker, M. (2021, May 12). Issues of trust in teachers' assessments: Can the profession move forward? *Impact: Journal of the Chartered College of Teaching, 12*.

Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education, 58 (Supplement C)*, 149–162.

Appendix

Specific subjects that teachers reported and the subject categories they were organised into.

Subject group	Subjects reported by teachers
Arts / Performing Arts	Music, Drama
Computer Science / Information and Communication Technology	Computing, Computer Science, Information and Communications Technology, Certificate in Digital Applications
Cross disciplinary	Global Perspectives, Critical Thinking, Personal Social Health and Economic education, Extended Project Qualification
English	English, International English Language Testing System, English Literature, English Language, English Language and Literature, English for Business and Finance, Writing, Teaching English as a Foreign Language, Cambridge A2 Key English Test and B1 Preliminary English Test, Functional English, Foundation English, First Language English, Second Language English, Linguistics
Humanities	History, Classics, Geography, Geology
Languages	Latin, Spanish, French, Italian, Afrikaans First / Additional Language, German
Mathematics	Mathematics, Functional Mathematics
Science	Science, Physics, Chemistry, Biology, Physical Education, Sport Science, Health Education: Food and Nutrition
Social Sciences	Psychology, Sociology, Criminology, Law, Business / Business Studies, Economics, Travel and Tourism, Religious Studies / Education, Divinity, Islamic Studies, Health and Social Care
Vocational	Teacher training, Leadership, Work Skills, Education and Training

Research News

Lisa Bowett and Jackie White (Research Division)

The following reports and articles have been published since *Research Matters*, Issue 40:

Journal articles and other publications

Kreijkes, P., Kewenig, V., Kuvalja, M., Lee, M., Hofman, J. M., Vitello, S., Sellen, A., Rintel, S., Goldstein, D. G., Rothschild, D., Tankelevitch, L., & Oates, T. (2026). Effects of LLM use and note-taking on reading comprehension and memory: A randomised experiment in secondary schools. *Computers & Education, 243,* 105514.

Cambridge University Press & Assessment contributed to a project in collaboration with the Centre for Education Systems, funded by the Nuffield Foundation. The Cambridge staff involved were:

Aramouni, N., Bowett, L., Brady, J., Ireland, J., Jin, C., Lieber, E., Majewska, D., Morrish, D., Nelson, S., Rushton, N., & Walland, E.

See here for outcomes of the work: https://www.ces.partners/curriculum-summary.

Research and statistics reports on our website

Benton, T., & Carroll, M. (2025). The biggest equating study in the world... ever.

Gill, T. (2025). The impact of the failure to account for KS2 cohort effects on awarding at GCSE.

Ireland, J., & Lestari, S. (2025). Digital divide: A literature review.

Majewska, D., Kreijkes, P., Greatorex, J., Jin, C., Saraoru, A. R., & Bowett, L. (2025). Learning theories resource.

Conference presentations

Suto, I., & Lim, C. H. J. (2025, December 6–10). *Empowering learners through wellbeing and digital awareness* [Paper presentation]. Cambridge Schools Conference, Bali, Indonesia.

The 2025 conference of the Association for Educational Assessment – Europe (AEA-Europe) took place in The Netherlands from 5 to 8 November, https://2025.aea-europe.net/. Our researchers presented five papers:

Constantinou, F., Miranda, A., & Pirola, S. (2025). *Understanding the construct of 'future skills': Some critical reflections emerging from a comparison of recent and older future skills frameworks* [Paper presentation].

Miranda, A., & Benton, T. (2025). *Utility of expert judgement for setting grade boundaries in England* [Paper presentation].

Vidal Rodeiro, C., & Gill, T. (2025). *Predictive validity of digital mock assessments* [Paper presentation].

Vidal Rodeiro, C., & Lim, C. H. J. (2025). *Assessment mode effects and their relationship with item characteristics* [Paper presentation].

Walland, E., & Rodrigues, A. (2025). *The validity of virtual labs for assessing science practical skills* [Paper presentation].

Blogs and podcasts

Constantinou, F., & Morrish, D. *Future skills: How they have evolved and why it matters* (28 January)

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

Research and statistics reports

Conference papers

Blogs

Research Matters (in full and as PDFs of individual articles)

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

Contents / Issue 41 / April 2026

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

CAMBRIDGE

The Assessment Network

Assessment Horizons Conference 2026

Join us in Cambridge or online on 23–24 April for a two-day conference bringing together members of the global assessment community to share insights, research, and best practice.

Be inspired by expert keynotes and stay up to date on the latest developments in areas such as assessment and AI.

For more details visit
https://www.cambridgeassessment.org.uk/the-network/assessment-horizons-conference