RESEARCH AND ANALYSIS

On-screen assessment and mode effects

A review of the effects of on-screen assessment on levels of engagement, cognitive demands and performance

Authors

- Harvey Dodds
- Yasmine El Masri
- Tim Stratton

With thanks to

- Zeek Sweiry
- Tom Bramley

How to cite this report:

Dodds, H. J., El Masri, Y., Stratton, T. (2025). *On-screen Assessment and Mode Effects: A review of the effects of on-screen assessment on levels of engagement, cognitive demands and performance*. (Ofqual research report 25/7280). Ofqual.
Retrieved from: https://www.gov.uk/government/publications/on-screen-assessment-and-mode-effects

Contents

Executive summary

Ofqual has conducted a wide-ranging research project into the risks, challenges, and benefits of using technology in assessment. Part of this research was concerned with the impact any move towards on-screen assessment (OSA) may have on students. Alongside primary research into the perceptions that students, teachers, and parents and guardians have of OSA, we undertook desk-based research to identify the evidence of the potential benefits and drawbacks of on-screen assessment. This review is one of the products of this desk-based research, and focuses on the effects of OSA on students' engagement with, cognitive demands during, and performance in assessment.

As the purpose of this review is to use the evidence in the literature to inform thinking about OSAs in England, it is worth highlighting from the outset that the literature is limited in several significant ways. Most of the field's findings are derived from experimental methods, often at a university level and often in a North American context. There are 94 sources cited in the review, of which 78 are empirical studies. Only 10 derived their findings from live summative assessments. Only 4 use data from high-stakes, summative assessments targeted at age groups comparable to those taking GCSEs or A Levels. Moreover, the literature spans a large period of time where technology, patterns of use, and access to technology have all changed and developed considerably. Covering such a broad period is unavoidable, as it is only possible to identify what is an intrinsic difference between paper and digital media and what is a contextual mode difference by taking such a long view. However, doing so brings its own challenges, and contributes to an overall need to consider the extent to which any given finding is likely to be relevant for OSAs in England.

Mode effects, affordances, and on-screen assessments

The term mode effect describes a situation where students perform differently on comparable tests that are taken via different media, such as a computer or pen and paper. A prominent cause of mode effects is a difference in the affordances – the possible actions a person believes can be taken while using a specific object – of the different testing modes.

OSAs can be defined as digital assessments that are taken by interacting with screen media. This will likely – although not always – involve using a combination of a display screen, keyboard, and some form of input to navigate the display (such as a mouse or touchscreen). Paper based assessments (PBAs) can be defined as assessments that are taken by interacting with a physical answer booklet with the use of a pen or pencil. The intrinsic differences of these media – alongside assumptions about how they can be used – means that the potential affordances of both are vastly different, and this provides the basis for mode effects.

The assessment mode can be thought of as having a direct effect on performance that is mediated by characteristics of the assessment user. It is important to investigate how the effect differs across modes, and if the effect of the same mode is different for different students.

Student preferences and engagement

The literature suggests that students generally prefer OSAs. A range of reasons for this preference are proposed, including a preference for typing and finding OSAs to be less stressful. In particular, different ways of doing assessments that are made easier by technology – such as gamification, adaptive testing and flexible on-demand testing – are often linked with student preference and/or engagement for OSA.

Nevertheless, the evidence is not unanimous in favour of OSAs, and the applicability of these elements in an English exams context is unclear. For example, it is unclear to what extent any increased preference and/or engagement is a result of the relative novelty of OSAs, and if levels of preference and/or engagement will decrease as OSAs become more widespread. Additionally, even if increased levels of preference and/or engagement are not the product of novelty, it is not clear that they will lead to better performance.

There are some important caveats that should also be borne in mind. Firstly, mirroring research undertaken by Ofqual, evidence in the literature highlights that students still value handwriting in some contexts. Secondly, negative experiences with OSAs can lead to reduced effort – this could be because of new distractions that result from the inherently different nature of the medium (such as the sound of a keyboard), or because of poor assessment design. Finally, research into engagement tends to be more common in relation to formative assessment due to its inherently lower stakes, and the literature on OSA is no exception. As such, the extent to which the evidence in the literature is relevant for high-stakes summative assessments is unclear.

Cognitive demands

Previous research indicates that screen media does pose more cognitive demands than paper. This is sometimes because OSAs are poorly designed, meaning these extra cognitive demands can be minimised by improving the design. Additionally, the cognitive demands of OSA can also be mitigated and managed if they arise from a lack of familiarity with the medium or the platform, as interventions that familiarise students with digital tools and/or provide digital upskilling can rectify this. However, the extra cognitive demands often come from the affordances that emerge through interacting with screen media (such as typing and scrolling). In particular, reading seems to be more demanding on screen than on paper. Some of these demands may come from a “dual task effect” where users of information and communications technology (ICT) must navigate the demands of the task at hand alongside the demands of interacting with the ICT. It is possible that the size of this effect has reduced in recent years as screen media has become more widespread (Eyre and colleagues, 2017). However, given the prevalence of touchscreen devices in the present day, it is important to consider if a dual task effect may still be prevalent for students who have to type assessments using a keyboard, which they may be less familiar with.

There does appear to be evidence throughout the literature that suggests that speededness may play a key role in mode effects: specifically, non-speeded objective tests do not appear to be liable to substantial mode effects whereas speeded tests are liable to substantial mode effects. Wheadon (2007) identified mode effects in GCSE biology and physics assessments that were offered on screen historically where students of middling-ability performed better in the OSA than the PBA. He argued this was due to the PBA being speeded and the OSA not being speeded. As such, dual administration of paper and digital assessments that are assumed to be equivalent may pose an issue as the cognitive demands of PBA may be higher.

While much of the evidence indicates that OSAs may pose greater cognitive demands than PBAs, there is some evidence that highlights that innovative OSAs may address excessive cognitive demands by reducing test anxiety. Even in this regard, though, the evidence is mixed, suggesting that any reduction in test anxiety is not an inherent property of OSAs, but rather that result of appropriately designed and deployed OSAs in a given circumstance.

Performance

Mode effects can occur for a variety of reasons. A useful way to categorise these reasons are differences in performance that appear to be related to student characteristics (namely age, gender, language, immigration status, ethnicity, socio-economic status, and underlying ability) and differences in performance that appear to be related to assessment characteristics (such as subject, item type, and so on). A small amount of literature (for example Jerrim 2016) considers the interaction of student and assessment characteristics in explaining performance differences, but this is currently an under-researched area.

Student characteristics

There are mixed findings as to whether performance on OSA differs by student characteristics. This is true when all students are considered, and when different subgroups of students are examined. A possible reason for this is that the relationships under consideration – that is, between categories (such as gender and age) and performance in OSA versus PBA – are unlikely to be fixed. Any performance difference is not likely to be an essential characteristic of, for example, gender. Instead, performance differences are likely to be a product of the assessment context (which includes, but is not limited to, characteristics of the student taking the assessment, the assessment subject, and the broader social and national context). Relatedly, it is also unclear to what extent some of these effects are independent of one another: immigration status and socioeconomic status are likely to be linked, and socioeconomic status may be linked to computer familiarity, again underscoring the importance of the broader context in understanding mode effects.

As a result of the fluidity of these relationships, the evidence about the direction and size of mode effects, as well as the underlying mechanism causing the mode effects, in the literature is also mixed. That is to say, there is evidence that OSAs can both decrease the performance gap between students of different socioeconomic status, genders, and underlying abilities, as well as evidence that it can increase the gap between the same subgroups. Similarly, there is no clear evidence as to why the gap narrows or widens: for example, some studies suggest it narrows because the bottom group is improving, whereas others suggest it narrows because the top group performs worse.

Assessment characteristics

The size and direction of mode effects appear to vary by subject, with better performance being associated with PBA in some subjects and with OSAs in other subjects. In particular, student performance is often considerably worse in maths assessments when they are taken on screen. This is possibly because mode effects are likely to be most pronounced when the capabilities and affordances of paper and screen are drastically different. Specifically, a range of tasks (such as jotting down, drawing shapes, and writing equations) can be experienced as being easier on paper because of intrinsic differences between the media. It is also important to consider that mode effects may vary within subjects, too – some areas of maths, such as graphing and geometry, are particularly prone to mode effects, for example.

There is mixed evidence for mode effects in constructed response items. This is despite evidence suggesting certain qualities of typed constructed responses tend to be stronger than handwritten responses. For example, it seems that typed responses are longer, contain better grammar and fewer spelling errors, and more relevant information.

Conclusion

Students appear to prefer OSA, and OSA may be more cognitively demanding than PBA. However, it is unclear how these 2 facts relate to performance in assessment. Moreover, there are inconsistent patterns of mode effects reported in the literature. This is further complicated by the geographic and temporal spread of the literature. It seems probable that any move to OSA in England will be associated with mode effects, but – given the mixed findings in the literature – anticipating what the size and direction of these mode effects can be difficult. As such, it is important to carefully and critically assess the relevance of any given finding before using it to inform the adoption and use of OSAs in England. Any given study is likely to have valuable insights that can inform practice, but it will require work to identify precisely what is relevant and what is not.

Introduction

Digital tools and computer devices are being increasingly used in education, be it for teaching and learning or for assessment purposes. Some jurisdictions and organisations have taken steps to transition their high-stakes assessments from the pen-and-paper mode to the computer screen. Key drivers include a willingness to take advantage of the opportunities that technology can offer to improve students' experiences of learning and assessment as well as a pressing need to prepare the young for the workplace which would eventually require them to operate confidently in a digitised environment (for example, see Ofqual, 2025 On-screen assessment research study).

England, like many other jurisdictions, has been exploring the benefits and challenges of a greater adoption of on-screen assessment (OSA) in its high-stakes sessional qualifications. In 2020, England's Office of Qualifications and Examinations Regulation (Ofqual) published a report highlighting the opportunities and barriers of moving high-stakes sessional assessments on screen in an effort to encourage the conversation. Incidentally, the COVID-19 pandemic made this step only more relevant with claims that an increased use of OSA could bolster the resilience of the sector in case of a large-scale disruption in the future. In 2021, Ofqual launched a large programme of work to examine the opportunities, challenges and risks of a potential increase in the use of OSAs in high-stakes qualifications in England. The programme aims to gather the necessary information in this area to ensure that future regulation and policy are grounded in evidence.

One strand of the evidence that Ofqual has been gathering relates to understanding the potential impacts of a shift in assessment mode on students. This is because polling research, commissioned by Ofqual (2023) into the perceptions that students and parents have towards on-screen assessment, indicated a majority of both groups believed students would find tasks associated with reading and writing on the computer – such as reading, scrolling, and typing – easy. Both groups tended to believe that mathematical tasks – such as typing equations, plotting graphs, and drawing geometric shapes – would be harder. More parents believed their children would finding typing equations (45%) and plotting graphs (42%) easy than hard (30% and 28%, respectively). It is therefore unsurprising that, in this same polling data, beliefs about the suitability of OSAs were largely subject dependent.

Research in the literature suggests that students interact differently with material presented to them on paper compared with material presented on screen, and in an assessment situation, this could lead to students performing differently depending on which assessment they take. These differences are known as "mode effects". To add complexity, these potential effects may be inconsistent across different subgroups of students (gender[1], socioeconomic status, ethnicity, ability, special educational needs, and so on).

We therefore sought to identify and review specific evidence about how completing different assessment tasks – such as reading, typing, scrolling, and completing mathematics tasks – may differ across mode, and how these differences between modes may themselves differ between students. In particular, this review considers how the mode through which students are being assessed (paper or screen) can positively or negatively impact:

1. their levels of engagement with the assessment
2. the cognitive demands they can experience during the assessment
3. their performance on the assessment

We have also undertaken another review (Ofqual, 2025) entitled *Making sense of mode effects* that cites many of the same findings covered in this review. However, while this review focuses on student performance, *Making sense of mode effects* is concerned with the implications of mode effects for assessment design.

Mode effects and affordances

A mode effect is when there is non-equivalence in psychometric item and/or test properties between different testing media (Buerger and colleagues, 2019). In plainer terms, the term mode effect describes a situation where students perform differently on comparable tests that are taken via different media, such as a computer and pen-and-paper. This non-equivalence may arise for any number of reasons – one example could be that some students may be more familiar with, and proficient in using, digital devices than others. That being said, differences in performance may arise in many instances because the affordances of the testing modes differ.

In the context of design, affordances are a way to describe the relationship between a person and an object: more specifically, they describe the possible actions a person believes can be taken by using an object (Norman, 2002). Affordances are neither positive or negative, nor are they solely an inherent property of a given mode. Instead, they are relational, with what a person believes they can achieve with an object key to determining whether it is an affordance in any given instance. The inherent differences of different media can be said to produce potential affordances,

[1] The term gender is used here to reflect the language used in the literature. In future primary research undertaken by Ofqual, we will distinguish between sex (meaning biological sex) and gender (meaning gender identity) in recognition of the two being distinct categories which, although overlapping, may have independent effects. In the studies cited in this review, it is rarely – if ever – made clear what is meant by gender, although it seems that gender is generally used as a synonym for sex.

but it is the input of a person that determines whether a potential affordance becomes a realised, actual affordance. For example, Norman (2002) highlights that the display, keyboard, pointing device, and selection buttons of a computer afford pointing, touching, looking, and clicking on every pixel of the display screen, but most of these actions are of little, or no, value. It is only when certain pixels are selected that meaningful actions – such as opening and closing a document – can be taken. As such, there is a difference between the potential affordances that arise when interacting with a computer, and the affordances that a computer user might perceive. Similarly, every millimetre of a piece of paper could be written on with a pen, but the action that a writer wishes to perform plus writing conventions (for example, left-to-right and top-to-bottom) and features such as lines or response boxes on the page may influence what they perceive as an affordance of the medium.

On-screen assessments (OSAs) can be defined as digital assessments that are taken by interacting with screen media. This will likely – although not always – involve using a combination of a display screen, keyboard, and some form of input to navigate the display (such as a mouse or touchscreen). Paper based assessments (PBAs) can be defined as assessments that are taken by interacting with a physical answer booklet with the use of a pen or pencil. The intrinsic differences of these media – alongside assumptions about how they can be used means that the potential affordances of both are very different, and this provides the basis for mode effects.

As such, the assessment mode can be thought of as having a direct effect on performance that is mediated by characteristics of the assessment user. Whether an assessment is oral, practical, on paper, or on screen, the mode of the assessment will therefore influence performance – the question is how the effect differs across modes, and if these different effects unfairly disadvantage some students. This review is specifically concerned with what effect OSA may have on performance when compared with PBA, and if these effects vary as a result of characteristics of the test-taker such as their age, gender, and underlying ability.

Literature at a glance

Tables 1, 2, and 3 (Appendix 1) list and categorise all the literature cited in this review that makes an empirical claim about mode effects, sorted by the educational level and type of assessment used.

In total, there are 94 sources cited from a wide range of journals and grey literature relevant to assessment, 16 of which are reviews or meta-analyses. The journals in which the studies are published are not always high impact, but all are either published by well-established publishers or included in well-established indexes. Of the 78 empirical studies, 35 generated evidence for their claims from samples at a higher education level, and 10 of the 78 used data from live summative assessments. Of the cited studies, only 4 – Backes & Cowan (2024), Keng and colleagues, (2008), King and colleagues (2011), and Wheadon (2007) - used data from high-stakes summative assessments targeted at age groups comparable to those taking GCSEs and A Levels. This is not unusual – in a conference presentation in 2014, Singer (2014, in Singer and Alexander, 2017A) reported that more than 75% of literature they surveyed pertained to reading comprehension involving undergraduate readers. There is therefore a clear gap in the literature, as there appear to be very few studies that have directly explored the impact of computer-based testing on high-stakes summative assessment for pre-18 education. The sources cited are from a range of international contexts, and subsequently a range of international school and educational systems. Where possible, for ease of interpretation, the age ranges of the samples used in research are grouped by the English equivalent key stage (KS): students between the ages of 5 and 10 can be thought of as being of KS1 and KS2 age; students between the ages of 11 and 14 can be thought of as being KS3 age; and students aged 15 and 16 can be thought of as being KS4 age.

Contradictory findings permeate the literature, making it difficult to any draw firm conclusions about the magnitude, direction and consistency of mode effects. The capabilities of ICT, student familiarity with ICT, the integration of ICT into learning and the classroom, and attitudes towards ICT might be completely different from one study to another, making their results difficult to compare. The different education levels and ages of the students, subjects, item types, and assessment stakes represent other factors that vary across studies, and which must be considered. Jerrim (2016) points to the importance of contextually situating findings, as he found that the strength and direction of different mode effects in his subset of PISA 2012 data were inconsistent across 32 participating countries, suggesting that broader contextual factors can affect students' performance on screen. Additionally, the literature included in this review covers more than 3 decades from 1993 until the present day. While, in the broadest terms, OSA refers to assessments taken on digital devices that use a screen as part of the human-computer interface, what OSA meant in practice in 1993 is likely very different from what it means in practice in 2025. Further, even in the present day, OSA is extremely diverse meaning the affordances differ from one context to the next. Nevertheless, a preponderance of studies that are one-off experiments undertaken at the higher education level (often in the United States) is a common theme across these 3 decades, as are conflicting results between studies. It is therefore important to avoid uncritically applying the results of a given study to the contemporary English education and assessment context and to appropriately weight the available evidence. This is because the diverse findings of the literature seem to suggest that differences in student performance across different modes may be the product of the interaction of different factors in a given assessment context (which would include factors such as student characteristics, assessment characteristics, and factors from the broader social and national context such as ICT proficiency and availability).

Student preferences and engagement

Some authors have argued that OSAs, especially those using innovative approaches to assessment, lead to higher levels of student engagement (for example, Bryant, 2017; Taylor, 2005) and may be the preferred mode of assessment among new generations of students (Taylor, 2005).

Studies conducted in the context of formative assessment, generally with higher education students across different geographical locations (such as South Africa and North America), point to various reasons behind students' preference for OSAs. These have included the convenience of assessment mode (Franic and colleagues, 2021), the ability to receive immediate feedback (Engelbrecht & Harding, 2004; Franic and colleagues, 2021), greater flexibility in administration time and location (Engelbrecht & Harding, 2004) and a less stressful testing experience (Engelbrecht & Harding, 2004). In a study conducted in Greece, Terzis and Economides (2011) identified a range of factors that could increase engagement with an OSA, with how enjoyable the OSA is to use, alongside clear and relevant content, being important for both genders. Ease of use was also important for females. While this was undertaken in the context of formative assessment, they offer potentially helpful insights into what may drive preferences.

It is, however, unclear whether these elements shape the views and preferences of secondary school students in England preparing for high-stakes examinations. A recent survey by Ofqual (2023) suggested a more mixed picture among students who were preparing for GCSE and A level qualifications and were asked questions about a hypothetical OSA. These students believed that some subjects (such as English language, computer science, history, geography, and business studies) are more suited to OSA than other subjects (such as mathematics, drama, art, PE, and music).

In the following subsections, evidence about student preferences and engagement is considered. It appears that students often report preferring OSAs for a range of reasons, including a preference for typing. The literature also suggests that computer familiarity and the degree to which the assessment is innovative[2] can be important.

[2] In this context, an assessment may be considered innovative if it utilises the capability of ICT to assess skills and knowledge in unconventional and/or creative ways. Innovative assessments can be directly contrasted with "paper-behind-glass" assessments, where the capacity to adapt the functionality or methods of assessment with ICT are not utilised, making the OSA comparable to a

However, it is unclear if increased preference is a result of the relative novelty of OSA, or if preference for OSAs is directly associated with better performance. Additionally, as much of the research into preference and engagement has been conducted in the context of formative assessment, the extent to which the following findings will hold translate to the context of high-stakes assessment in England is unclear.

Computer familiarity

Advocates for OSA have suggested that new generations of students are more familiar with the screen mode than the paper mode (for example, Taylor, 2005). Prensky (2001a; 2001b) suggested that younger generations have grown up around computer devices as “digital natives” with frequent interactions with various digital technologies. The idea that there are sweeping differences between so-called “digital natives” and “digital immigrants” is contested (see, for example, Heslper & Eynon, 2010; Selwyn, 2009). However, Prensky’s (2001a; 2001b) description of the ubiquity of technology in the lives of younger people does mirror figures reported by Ofcom (2024). In this report, which detailed the widespread access young people now have to both hardware and the internet, it was noted that 96% of children accessed the internet in 2023 (including the low proportion of 3 and 4 year olds who do not go online), and over 90% of children aged 11 plus owned their own phone which they would use to access the internet.

Widespread screen use among young people does not only mean they can interact with digital devices with familiarity, but that their interactions with technology may also be shaping their cognitive processes making them operate digital devices with greater ease. Conversely, a study by Hughes and colleagues (2010) found that secondary school students preferred completing their mathematics GCSE examinations on paper as they found it more “familiar” and “natural”. Hughes and colleagues do not explain exactly why the students felt this way, although the study’s other findings indicate that the ability to freely annotate and jot down workings on paper may play a role. Nevertheless, their findings underscore the important role that familiarity can play in assessment preferences. The importance of familiarity was also clear in polling commissioned by Ofqual into perceptions of OSAs, with more than three-fifths of students indicating they would be concerned if they did not know what an exam looked like prior to taking it or if they were unable to practice taking OSAs in school prior to the exam itself (Ofqual, 2023).

PBA other than the mode of input. In terms of adapting a PBA into an OSA, an innovative assessment can be thought of as a transposition, whereas a paper-behind-glass assessment would be a direct translation.

Typing versus handwriting

Students' preference for the screen versus the paper mode of assessment has been associated with their preference for typing over handwriting (for example, Whithaus and colleagues, 2008). Some researchers have related this preference to students' levels of familiarity with computers and argued it would be unacceptable to expect new generations of students to handwrite an exam (Taylor, 2005). If true, this concern would be particularly relevant to the examination context in general qualifications in England where most students are expected to handwrite responses to a large number of open-ended questions to which they must construct a response (henceforth constructed response items), including essay questions.

Research in higher education has considered the advantages and disadvantages of typed versus handwritten examinations (Chan, 2023). Chan's systematic review suggests that handwritten examinations are associated with lower legibility of students' scripts and feelings of discomfort which are claimed to be detrimental to students' performance.

Nevertheless, it is important to emphasise that any preference for typing is not fixed and universal. Not all school-aged students have the same access to ICT devices – a third may not have access to appropriate technology for learning at home (Ofcom, 2024). Therefore, if preference for typing is a function of computer familiarity, writing preference could vary accordingly. Further, evidence from the past decade indicates that students believe that handwriting still has strengths and virtues and that preference for typing or handwriting is not fixed, but situation-specific (Farinosi and colleagues, 2016; Fortunati & Vincent, 2013).

Innovative assessments

OSAs have enabled innovative formats of assessment including computer adaptive tests, gamification, and technology enhanced items that are not possible on paper. Computer adaptive tests have a large bank of items and tailor the level of difficulty of questions presented to test-takers based on their prior answers to accurately determine their ability. Gamification is where a test is designed to include aspects of game-playing. In an on-screen context, this can even include tests in the style of video games. Technology enhanced items are items that are enhanced by the affordances of information and communication technology (ICT) such as multimedia stimuli provided with the question prompt, the format of the response (for example, the drag and drop function), and the way in which an examinee responds to the item (for example, using the microphone to record an audio response).

There are claims that innovative assessments generate higher levels of engagement among students (Bryant, 2017; International Test Commission and Association of Test Publishers (ITC & ATP), 2022; Taylor, 2005). In particular, computer adaptive tests and gamified tests can do so by matching the test to the test-taker's ability (ITC & ATP, 2022). However, it should be stressed that greater engagement as a result of adaptivity may not automatically equate to better performance: Ling and colleagues (2017) conducted an experiment with 789 KS3-aged students comparing computer adaptive tests to fixed OSAs and found that higher engagement with a computer adaptive test was not associated with better performance when compared to the fixed assessment.

It is important to consider whether increased student engagement with, and preference for, OSA is inherent to the mode of assessment or the result of its novelty. Evidence in the literature is mixed. On the one hand, Dahan Golan and colleagues (2018) and Halamish and Elbaz (2020) found that, among primary school students in Israel, preference for reading on screen decreased after students had completed some screen reading and an on-screen comprehension test, suggesting that preference and engagement with OSA decreased when the novelty of the assessment mode wore off. On the other hand, Fluck and colleagues (2009) found a significant positive correlation between prior use of, and preference for, OSA among Australian students in higher education, which suggests that novelty does not always explain preference. Similarly, other research at a higher education level has found that increased exposure may be associated with greater acceptance of OSA (Boevé and colleagues, 2015).

While innovative OSAs can have a range of positive impacts on students' assessment experience, evidence from learning contexts highlights that innovative technology can sometimes have a detrimental effect on the quality of students' performance. For example, Makranksy and colleagues (2019) found that immersive virtual reality could increase engagement among students in higher education, but that it also led to less learning and higher cognitive load. If these findings translate to assessment, it may suggest that increased engagement could come at the expense of poorer performance because the immersion may overload and/or distract students.

Negative experiences with OSA

There is also some research that outlines negative experiences with OSA. In a Swedish higher education context, students felt increased stress and fatigue with OSA (Wästlund and colleagues, 2005). The mode can also introduce new distractions, such as the noise of the keyboard (Fluck and colleagues, 2009; Lim and colleagues, 2006) and difficulties with maintaining examination conditions (Jerrim, 2016). Additionally, Fluck and colleagues (2009) reported that student experience was negatively impacted by the design of the assessment in an OSA undertaken in an Australian higher education context. Specifically, the students were required to switch between multiple windows which caused frustration.

It should be noted that all the previously cited studies detailing negative experiences with OSA are from 2012 or earlier. It might be possible that, as the technical capabilities of digital devices have improved and as information technology use has become more widespread, some of the issues have become less prominent. Further research from this time period suggests OSAs are not inherently positive or negative, but that students' experiences are contingent on the nature of the OSA itself. For example, Ricketts and Wilks (2002) found when they optimised the presentation of the test content for screen viewing, students overwhelmingly preferred OSAs. This would suggest the quality of the test design can be key to how well students experience OSAs. Additionally, Lim and colleagues (2006) found that nearly 80% of their sample of 114 medical students at the National University of Singapore preferred OSA for a multiple-choice exam, but this reduced to 54.4% for a modified essay question. As with preference for typing versus handwriting, this could suggest that preference for OSA should not be thought of as fixed, but as situation-specific.

Cognitive demands

The intrinsic differences of digital and pen-and-paper assessment modes means that interacting with these formats can be regarded as different tasks. For example, when reading on paper, text is in a fixed location and writing with pen and paper requires manually manipulating physical implements. In contrast, reading on a computer screen often involves scrolling, meaning text is comparatively fluid, while writing requires the use of a keyboard to type words on a virtual page. These tasks will be even more different on tablet and mobile devices.

These intrinsic differences may pose different cognitive demands (see for example, Garland & Noyes, 2004), some of which may be irrelevant to the construct being measured, or be relatively more (or less) cognitively demanding (for example, Noyes & Garland, 2003). Much of the literature in this area is drawn from research on the assessment of reading comprehension, with screen reading appearing to be both a qualitatively different, and more cognitively demanding, task to reading on paper. Although focused on a specific area of assessment, it is pertinent for assessment more broadly as reading and understanding the questions in an assessment is a pre-requisite for successfully completing the assessment. This literature is covered in the following sections, alongside literature that considers the effect of other factors on cognitive demands such as typing, the design of the OSA, and the hardware involved. Literature that considers the cognitive processes that may lead students to approach, or navigate, OSAs differently to PBAs is also considered, as is literature arguing that OSA may have a beneficial impact on cognitive demands.

Overall, existing research seems to suggest that screen media poses more cognitive demands than paper, and that some of these additional cognitive demands may come from the affordances – such as typing and scrolling – that emerge through interacting with screen media. Other additional cognitive demands may be a result of contextual factors, and may therefore be managed or mitigated. Some research also indicates that OSA may help to reduce additional cognitive demands by reducing test anxiety, although there is mixed evidence for this. As with engagement and preference, it is unclear whether, and to what extent, the greater cognitive demands associated with OSA impact performance.

Reading comprehension

On-screen reading has frequently been associated with poorer comprehension at all education levels, suggesting that a mode effect exists (see, for example, Eyre and colleagues, 2017; Halamish & Elbaz, 2020; Mangen, and colleagues, 2013; Singer & Alexander, 2017A; Støle and colleagues, 2020; Wästlund and colleagues, 2005). The effect size is usually small, but this consistent finding has been further confirmed in meta-analyses and literature reviews.

Research suggests that text may be cognitively processed differently when it is read from a screen compared with when it is read from paper (Garland & Noyes, 2004; Mangen, 2008; Mangen, 2010 in Jensen, 2020; Noyes & Garland, 2003; Singer Trakhman and colleagues 2017). In particular, it appears as though screen-reading may lend itself more to "shallow"[3] reading (Mangen, 2008). Noyes and Garland (2003) note that Tulving (1985) makes a distinction between 2 retrieval processes: "Remember" and "Know", with the former describing information that is consciously recalled with contextual information about the way in which the information was learned, and the latter describing an ingrained, more unconscious familiarity with the information which can be more easily recalled. Noyes and Garland conducted an experiment where 50 university students with no prior knowledge read materials related to economics. The participants then completed comprehension questions and were asked to self-report how they retrieved the answer. They found that screen readers were twice as likely to "remember" information than "know" it, whereas levels

[3] The "shallow" reading Mangen refers to describes less focused. superficial scanning of text on screen where the reader is easily distracted due to cognitive exhaustion. This occurs because the intangibility of digital texts means that they are constantly changing, and this results in highly visually stimulating environment. Moreover, digital environments are also associated with an array of possible distractors (e.g. noise from mouse clicks, hyperlinks, and so on). In contrast, Mangen argues that deeper, more focused and immersive reading is intrinsic to the printed form, as it is a less visually stimulating environment and is shorn of the intrinsic distractions associated with screen media.

of “remember” and “know” were similar for paper readers. In discussing their results, Noyes and Garland (2003; Garland & Noyes, 2004) suggest that the visual variables associated with screens – such as screen refresh rates, contrast levels, and fluctuating luminance – may reduce working memory capacity, which ultimately limits the cognitive resources available for consolidating information in long-term memory. However, given the development of display technology since 2004, it is unclear if these factors continue to impact comprehension.

However, some studies have identified additional factors which moderate the relationship between mode and reading comprehension. For example, in a systematic literature review of 36 studies from 2001 to 2017, Singer and Alexander (2017B) found that there was no significant difference in comprehension between modes for texts under 500 words, but when texts were longer than 500 words, there was better comprehension on paper. Clinton (2019) found that comprehension was poorer on screen only for expository texts, with screen-reading having a negligible effect on narrative texts. This finding was echoed by Delgado and colleagues (2018), who also found an effect for texts which were a mixture of narrative and expository text. In a sample of between-participants studies included in their meta-analysis, Delgado and colleagues also found that the advantage of reading on paper rather than on screen increased year-on-year between 2000 and 2017, which was a significant effect and explained a large amount of the differences in results. This would suggest the negative effect of screen reading is increasing over time, despite the increased prominence of screen reading. In contrast, although Kong and colleagues (2018) found in their meta-analysis that comprehension was higher when reading from paper, they identified a trend that suggested a diminishing advantage of paper-based reading post-2013. Although this was not statistically significant at conventional levels, the authors did suggest that the trend could meaningfully inform the research questions of future research.

The effect of scrolling

Two key differences between reading on paper and on screen are that the paper is tangible and the location of the text on paper is fixed, while on-screen texts are intangible and typically have a fluid location requiring users to scroll up and down the page (Dillon, 1992). Støle and colleagues (2018) note that better spatial mental representations of the text layout (such as remembering where information was found or the order of different subheadings) is associated with better comprehension, and that the fixed nature of text on paper can provide spatial cues. In contrast, the fluidity of text on screen can mean these cues are absent. Sanchez and Wiley (2009) conducted a study with university students in the US and found that scrolling reduced their comprehension of complex topics on webpages. They further found that this effect was most pronounced for readers with lower working memory capacity. The authors outlined 3 possible explanations for further research to investigate. Firstly, they suggested that the need to maintain a surface representation of the text while comprehending the content could create additional cognitive load which especially affected readers with low working memory capacity. Secondly, they suggested that readers with low working memory capacity may struggle to control their attention when reading texts that required scrolling, and that this could lead to them becoming disoriented or lost during reading. Thirdly, they suggested the absence of page breaks may mean low working memory capacity readers are not prompted to summarise, integrate, and consolidate what has just been read, resulting in them trying to consolidate a longer text than their working memory capacity allows them to remember.

Some evidence suggests the cognitive demand of scrolling may impact assessment performance. Bridgeman and colleagues (2003) asked 357 grade 11 students who intended to continue into higher education to answer reading and writing and maths questions from US SAT exams[4] to explore how screen size, resolution, and display rate might affect performance. They found participants who used the 15-inch display reported that scrolling interfered with taking the test, and that smaller screens and resolutions were associated with poorer performance on the reading items. In contrast, there was no significant difference for the maths items where little scrolling was required for any of the conditions. Bridgeman and colleagues' (2003) findings were only partially mirrored in Keng and colleagues' (2008) study, which found mode effects due to scrolling in both reading and mathematics summative assessments in Texas. Students performed worse on screen in reading questions accompanied by longer passages.

Importantly, even though there is consistent evidence to suggest that comprehension of on-screen texts is poorer, there is also consistent evidence to suggest that readers are unaware of this. Calibration is a metacognitive skill that enables learners to accurately judge their own performance (Pieschl, 2009). On-screen reading has been associated with poor calibration, as readers misjudge and overestimate their level of understanding of an on-screen texts (Ackerman & Goldsmith, 2011; Halamish & Elbaz, 2020; Singer & Alexander, 2017A). This finding was supported in a meta-analysis of 12 studies from 2008 to 2018 conducted by Clinton (2019), who found calibration was better when reading on paper, albeit with a small effect size. It should be considered that, when summarising the literature, Singer and Alexander (2017A) note that numerous studies have identified poor calibration among learners across a range of educational tasks (including those completed on paper) and

[4] The Standardized Achievement Test (SAT) is an assessment offered by the College Board that is used for university admissions in the USA.

subject domains at different educational levels, suggesting that this is not problem exclusive to ICT.

Response strategies and ways of working

Studies identifying a mode effect in strategy use and ways of working have been mostly carried out in the context of answering mathematics examination questions (for example, Hughes and colleagues, 2010, Johnson & Green, 2006; Keng and colleagues, 2008; Logan, 2015; Threlfall and colleagues, 2007). This literature suggests that students tend to have a higher performance on questions they answer on paper partly because they are better able to use familiar strategies to solve the questions, such as annotating diagrams and graphs (Hughes and colleagues, 2010), using jotting paper (Hughes and colleagues, 2010; Keng and colleagues, 2008; Logan, 2015; Johnson & Green, 2006), rotating figures (Johnson & Green, 2006) and showing their working (Logan 2015; Threlfall and colleagues, 2007). In particular, students' ability to show their working more easily on paper means that, in multi-step questions, they are better able to co-ordinate and manage multiple pieces of information as well as monitor their thinking (Logan, 2015). It is worth noting though that Logan's study used tablets (iPads) as screen devices, so it is unclear whether the challenge of showing working is related to keyboard typing or 'writing' using a touchscreen. Also, somewhat counter-intuitively, Johnson and Green (2006) found that, in 9 of 16 items, students completing questions on screen were more likely to show their working than those completing them on paper, while the converse was only the case for 4 items. However, they also noted there was a greater likelihood of students showing working for more difficult items, irrespective of the mode.

While some studies suggest that students answering mathematics questions on paper have a more flexible approach to answer questions (Johnson & Green, 2006), Threlfall and colleagues (2007) found that in some items where performance was better on screen than on paper, students answering questions on screen were able to employ strategies not applicable on paper due to the affordance of the medium. In geometry for example, Threlfall and colleagues found that students could reduce the cognitive load associated with symmetry questions without having to infer or visualise the line of symmetry; they just needed to experiment with placement on screen until they could recognise symmetry. Similarly, on another item students needed to permutate 6 cards each with a digit from 1 to 6 to answer the question "324 = ___ - ___"[5]. On paper, students would need to calculate the answer before

[5] The correct answer is 324 = 645 - 321

committing it to paper, whereas the drag and drop functionality on screen meant that students could keep trying to move the cards around until they were happy with the equation, and they avoided the risk of using the same digit more than once. This had the effect of reducing the cognitive load of the task, leading to better performance on screen.[6]

Moving through parts of the assessment

The literature identifies 2 particularly salient ways in which the assessment mode can influence how a student works through an assessment. The first relates to item review. Students completing an assessment on paper are free to move back and forth through pages and can review their answers, but this is not always possible with OSAs depending on the platform design. Not being able to move seamlessly through parts of the assessments may impact performance (Johnson & Green, 2006; Revuelta and colleagues, 2003). In a review of the literature, Lynch (2022) notes that mode effect studies concerned with item review have found mixed results, but nevertheless recommends that OSAs should allow item review to minimise possible sources of mode effects.

The second relates to the issue of speededness. When summarising the literature, Wheadon (2007) identified 3 important recurring findings. Of interest here are 2 of the findings: firstly, that non-speeded objective tests are not liable to substantial mode effects whereas, secondly, speeded tests are liable to substantial mode effects. Consistent with Ofqual (He & El Masri, 2025) work on the topic of speededness in PBA, speededness in OSA seems to be linked to student ability. For example, Karay and colleagues (2015) found no significant difference in the overall score between modes, but that students doing OSA needed significantly less time to complete the test, and this was particularly true for the highest performing students. Wheadon (2007) reported the findings of a study into the comparability on on-screen and paper-based GCSE biology and physics foundation tier papers. He found that students of middling ability who sat the OSA performed better on items later in the test than students of the same ability who sat the PBA. To explain this finding, Wheadon suggests that the OSA took less time to complete, which removed the speeded pressure from items at the end of the test. It should be noted, though, that Keng and colleagues (2008) found no evidence of an item position effect, although they did not investigate the speededness of the test in their study. This may suggest

[6] These items are discussed further in Sweiry, E. (2025). *Making sense of mode effects: A framework for anticipating performance differences in equivalent paper-based and digital test items.* (Ofqual research report 25/7278). Ofqual.

that specific contextual factors – such as test difficulty, student ability, and how they interact to inform the speeded pressure of the test – make item position salient.

Ackerman and Goldsmith (2011) conducted research with undergraduates at an Israeli university. They ran 2 experiments, where participants studied 6 texts of 1,000 to 1,200 words, either on paper or on screen, before completing a multiple-choice test of comprehension and memory of details using the same medium. In the first experiment, participants were given a fixed amount of time to study and be tested on each text, and no mode effect was found. In the second experiment, participants were given 90 minutes in which to complete the tasks which they could use how they wished. In this self-regulated scenario, on-screen performance was poorer. As such, Ackerman and Goldsmith suggested that meta-cognitive issues can be a cause of mode effects, with the digital medium being associated with poorer self-regulation and self-awareness. Similarly, Delgado and colleagues (2018) found a small effect suggesting that the advantage of print reading was larger in time-constrained situations when compared with self-paced reading, although this small effect only explained 5% of the differences between results.

Other factors

For OSAs that require students to type in responses, research suggests that students who are not proficient typists are likely to find typing taxing and this additional demand may negatively impact on the quality of their typed script and subsequently their overall performance on the test (Connelly and colleagues, 2007). Typing or handwriting responses have also been associated with how easy it is to edit written work (for example, Ceka & O’Geen, 2019), the strategies used for text composition (for example, Lee, 2002), and the vocabulary used and writing style (for example, Charman, 2013; Mogey & Hartley, 2013; Whithaus and colleagues 2008). However, some of the evidence is conflicting, and Lee (2002) noted that the strategy utilised might ultimately vary from individual to individual.

While students may use various strategies to reduce the cognitive load associated with specific tasks or particular modes of assessments, test developers can minimise some demands that are linked to computer use. For instance, in an experiment with students from a Swedish university, Wästlund and colleagues (2008) found that optimising page layout for on-screen viewing by adapting page size to fit the screen and eliminating the need to scroll can decrease some of the issues with cognitive load usually associated with on-screen reading. Rickets and Wilks (2002) found scrolling affected performance among undergraduate biology students. However, performance in, and preference for, OSA both increased among the subsequent year’s cohort who undertook a re-designed assessment that did not require scrolling.

Similar to the findings of Wästlund and colleagues, this could suggest that removing scrolling may reduce some of the cognitive issues associated with OSAs.

Relatedly, Benedetto and colleagues (2013) found that, among a small sample of adults, optimising on-screen viewing by using E-ink displays (which are used by some popular e-readers such as the Kindle Paperwhite) can lead to similar levels of visual fatigue to that experienced when reading on paper, whereas liquid crystal displays (LCD) commonly used in devices such as smartphones, tablets, and laptops lead to higher levels of visual fatigue. Additionally, while Bridgeman and colleagues (2003) found that screen size could affect performance, King and colleagues (2011) found no impact when examining the use of netbook laptops (small, lightweight laptops primarily designed for easy access to the internet) in K-12 geography, geometry, and English assessments in Texas. Finally, Connelly and colleagues (2007) theorise that targeted interventions in the form of typing lessons may help to improve typing fluency and the overall quality of typed work, because typing will subsequently pose less of a cognitive demand.

The above examples would add weight to the possibility, as outlined by Eyre and colleagues (2017), that advances in technology and increased student familiarity with computers may be diminishing the “dual task effect”, which describes how ICT users must simultaneously navigate the cognitive demands of the task and the cognitive demands of interacting with the ICT. Therefore, while much of the literature does suggest that OSA may impose additional cognitive demands on examinees, these demands can sometimes be mitigated and are perhaps not enduring effects of the mode itself, but are contextually situated. Nevertheless, it is important to consider that, in the context of the mid-2020s, students and young people are likely to be significantly more familiar with touchscreen devices – particularly mobile phones – than they are with both desktop and laptop computers (see, for example, Ofcom, 2022; 2023; 2024). As such, there is a possibility of a “dual task effect” among young people who may take OSAs using a computer, but may be more familiar with other digital devices with different affordances. Future research into the shape and form of any “dual task effect” in the present day may therefore be beneficial.

There may be some ways in which OSA may beneficially address the cognitive demands of examinations. For example, test anxiety is associated with poorer cognitive performance in more complex tasks (Dutke & Stöber, 2001). Certain capabilities associated with OSAs may help to reduce test anxiety in some instances. For example, some studies (such as Mavridis & Tsiatsos, 2016; Turan & Meral, 2018) have found that gamified OSAs can be associated with lower test anxiety.[7] Another way in which OSA may reduce test anxiety is through the ways in which it can increase the control that test-takers perceive they have of the situation.

[7] Although by no means always – see Albuquerque and colleagues (2017), for example.

For example, in a literature review, Wise (2019) notes that self-adaptive tests – where the test-taker selects the difficulty of the next question, and which are easier to administer via computer – have positive effects on text anxiety, even compared with other forms of OSA. In another example, Cassady and colleagues (2001) undertook research with undergraduate psychology students from Ball State University taking a non-adaptive test. They found that there were no disadvantages associated with OSA in terms of anxiety and emotionality (a dimension of test anxiety) as measured through self-reports using the Cognitive Test Anxiety scale (Cassady & Johnson, 2002), and students who took the OSA perceived the test as less threatening compared with those who took the PBA, and the authors argued that test threat is conceptually related to anxiety. The students who took the PBA sat the test in a group-setting at a pre-assigned time, while the students who sat the OSA were able to do so by dropping-in to a secure computer-based testing laboratory at any point over the course of a week, which are significant confounds and prohibit the possibility of portraying this as a pure mode effect. However, flexible anytime high-stakes assessment is much more feasible with OSA than PBA, and it points towards a way in which increased perceived control may decrease test anxiety.

It is important to stress, though, that OSAs on their own cannot remove test anxiety, and indeed may worsen text anxiety, or even introduce new forms of anxiety specific to computer testing, and this can have negative consequences for performance. For example, Ortner and Caspers (2011) found that computer adaptive tests could have a negative impact on the performance of anxious test-takers because of the volume of questions they could get wrong as the test calibrated to their ability. Shermis and Lombard (1998) found that, with a sample of undergraduates from the United States, computer test anxiety in particular was associated with poorer performance in OSAs of reading and maths, although it did not impact performance on written essays in a statistically significant way.

Student performance

The mode of assessment may not only affect students' levels of engagement and cognitive demands placed on them, but – perhaps in part because of these other effects – may also impact how well students do in the assessment. It could be argued that if performance is affected, but the impact is evenly distributed across the ability spectrum, it will only be a concern if there are simultaneous on-screen and paper-based versions of the same assessment that need to be comparable. However, if the assessment mode affects the performance of different groups of students to different extents, it will be much more problematic.

This section examines the extent to which the mode of assessment affects students' performance. It is split into 2 subsections, with the first exploring the degree to which the mode effects are systematically related to identifiable person characteristics. Specifically, it considers the degree to which there appears to be mode effects when the entire student cohort is considered before the effect of different individual level characteristics – such as gender, ethnicity, and socioeconomic status – are considered. The second subsection explores the degree to which mode effects are moderated by features of the assessment itself, such as item type and subject. Overall, there does not seem to be consistent evidence suggesting that OSA is associated with either better or worse performance when compared to PBA. The size and direction of mode effects appear to vary across demographic sub-groups, subjects, and item types, and arguably this shows that performance differences are a product of a given assessment context (which includes the student, a given assessment, and macro-level factors in the broader context). This again highlights the difficulty of predicting the precise of impact of any move towards OSA based on the current literature.

Students' characteristics

All students

There is mixed evidence in the literature as to whether students perform better on screen or on paper. Although some studies have found no significant differences in performance across the 2 modes among children aged 5 to 10 (Eid, 2005) and 11 to 14 (Poggio and colleagues, 2005), as well as at K-12 (Wang, and colleagues, 2007[8]) and higher education levels (Akbay and colleagues, 2021), other studies examining the same range of age groups have found better performance on paper than on screen (Bennett and colleagues, 2008; Bridgeman and colleagues, 2003; Dahan Golan and colleagues, 2018; Mead & Drasgow, 1993). Backes and Cowan (2024) investigated mode effects following a transition in online testing in Massachusetts. Their research covers a seven-year period from 2012 to 2018 and compares the 3 different statewide standardised tests that were administered during that time: a paper only version of the Massachusetts Comprehensive Assessment System (MCAS) used until 2016; and assessment introduced by the Partnership for Assessment of Readiness for College and Careers (PARCC) used in 2015 and 2016; and a new version of the MCAS used from 2017 that was delivered on paper

[8] Bennet and colleagues (2008) highlight some important limitations of this meta-analysis, highlighting the need to treat its findings cautiously

and on screen initially before becoming an exclusively OSA from 2019. There was evidence of mode effects in the OSAs for both the PARCC and MCAS assessments, but the mode effect in the MCAS was substantially reduced through equating measures.

Of the studies cited in this subsection, most that found no significant differences were in mathematics, whereas most which identified better performance on paper were in assessments related to English. Nevertheless, it appears unlikely that the subject of the assessment can explain the mixed findings: Bennett and colleagues (2008) found students performed better on paper in maths, and other studies cited elsewhere in this review (for example, Hughes and colleagues; Johnson & Green, 2006; Threlfall and colleagues, 2007) have identified issues with OSAs in mathematics. The role of subject is considered more fully in a subsequent subsection.

Age

Threlfall and colleagues (2007) found that KS2 students performed 3% better on on-screen mathematics assessments while KS3 students performed 5% better on paper-based mathematics assessments. They describe this difference in performance as limited and suggest that the performance in the OSA and PBA was comparable. However, there are 2 important caveats to this. Firstly, it would be unlikely that differences of this magnitude would be considered tolerable in high stakes assessments in England. Secondly, Threlfall and colleagues stress that, even though they argue that the assessments are comparable, the aggregate level scores masked some larger differences at item level. In discussing these item level differences, Threlfall and colleagues focused on the properties of the items themselves and did not consider age as a factor. Liu and colleagues (2016) similarly found that, among items identified as having performance related mode effects in the 2015 PARCC assessment, in grades 3 to 8 more flagged items favoured OSA, while in high school more flagged items favoured PBA.

In contrast to the findings of Threlfall and colleagues and Liu and colleagues, Wu and colleagues (2015) found that older students performed better on technology enhanced items than younger students. Further, it seems possible that in the 17 years since Threlfall and colleagues' findings were published, and the decade since the PARCC assessment, the relationship between age and performance in OSA may have changed. In 2007, the greater familiarity that younger students may have had with technology compared to older students may have explained why age moderated performance. ICT is far more ubiquitous in the present day, and there is evidence to suggest that, as students progress through education, their ICT use increases (for example, Langer and colleagues, 2016). As such, it seems likely that the relationship between age and OSA performance, should one exist, may have changed over time.

Gender

Overall, the evidence is mixed on whether there are consistently observed group differences in performance between OSAs and PBA in relation to gender. Using data from an experiment which contrasted 3 testing modalities (one PBA and 2 OSA) among students in the USA from the equivalent ages of year 4 through to year 12, Hardcastle and colleagues (2017) found only one small gender difference (where KS3-aged males performed marginally better than females in one of the OSA testing modalities) in the performance of students who completed a science OSA. Jerrim (2016) found that, on average, 15-year-old males performed better in both PBA and OSA in mathematics in a subset of PISA 2012 data, but that this gender attainment gap was larger for OSA than PBA. However, the size of this difference varied from country to country. Jerrim draws attention to the example of the USA and Shanghai, where both countries had a gender difference of approximately 5 PISA points for PBAs. With OSA, Jerrim found that the gender gap was closed in the USA (difference of 0 points) while it increased to 20 points in Shanghai, which further strengthens the notion that performance differences are likely to be a product the interaction of student and assessment characteristics with macro-level factors in a given assessment context. Other studies have found no interaction between gender and testing mode (Horkay and colleagues, 2006; MacCann, 2005).

The greater gender gap associated with the on-screen mode in mathematics observed in Jerrim's analysis is consistent with findings in other studies. Gallagher and colleagues (2000) conducted secondary data analysis on operational testing data and data from experimental administrations of OSAs. This included 1,401 SAT entries from 1996; 2 samples who took the Graduate Management Admissions Test in 1996 and 1997 totalling 8,029 entries; 3,791 Test of English as a Foreign Language test-takers from 1997 to 1998; two samples of GRE General test-takers totalling 503,168 entries from 1996 to 1997; and 119,609 Praxis Professional Assessments for Beginning Teachers entries from 1993 to 1997. They identified that some ethnic groups appeared to benefit from OSA, and females appeared to perform worse in OSAs in some tests. When considering the joint effects of gender and ethnicity, they found that OSA had, on average, a small negative impact on the performance of white females (but no other groups) in items drawn across all domains from the GRE General Test, the Graduate Management Admissions Test, and the Praxis Professional Assessments for Beginning Teachers.

In a subset of 145,953 students from 26 countries who took the core PISA 2012 reading assessment and an additional on-screen reading assessment, Borgonovi (2016) found that males tended to perform worse in both OSA and PBA, but that the performance gap between the genders was smaller in OSA. Borgonovi's analysis also indicated that moderate use of single-player video games was associated with performance advantage, which they suggest may be because of spatial and navigational skills that video games help to develop, and which are required for screen, but not paper, reading. In contrast, greater use of collaborative online games was associated with worse performance (particularly for lower performing students and in the paper version of the test). As males were more likely to play video games, more likely to play collaborative online games, and more likely to play video games excessively, Borgonovi concludes that, in many countries, these patterns of game-usage are strongly associated with gender differences in performance in the reading assessment.

Earlier work often predicted that males would perform better in OSA due to innate preferences for certain media forms (for example Martin & Binkley, 2009). However, the essentialist gender assumptions underpinning some of this earlier work have become contested, and there are a range of studies which suggest that the real relationship between gender and assessment mode is more complex than innate gendered preferences for certain media. A recent study by Jensen (2020) in fact reported the opposite pattern in mode effects and gender performance. Data from a field study for the 2016 Norwegian National Reading Test suggested that female KS4-aged students had better comprehension on screen while males had higher comprehension on paper. Wagner and colleagues (2021) report the results of a field experiment that explored mode effects in the low stakes German VERA assessment in the domains of reading and orthography. Reflecting the streamed nature of German education, there are 2 test booklets: booklet 1 is recommended for schools with less-academically able students, and booklet 2 is recommended for schools with more-academically able students. In booklet 1, females performed significantly better than males in both domains, but this difference was significantly smaller in the OSA mode than in the PBA mode. In contrast, in booklet 2, females performed significantly better than males in both domains, but this difference was non-significantly larger in the OSA mode. When taken alongside the finding that the mode effect in the VERA assessment was larger among students with lower academic achievement than among those with higher academic achievement, it may be the case that the moderating effect of gender is itself moderated by underlying ability.

Language background, immigration status, and ethnicity

A number of studies have considered how factors such as language background, immigration status, and ethnicity may interact with assessment mode to impact performance. It is important to stress that, although these factors may be linked, they are nevertheless discrete and do not always co-occur. As such, a mode effect, for example, among English as an additional language speakers should not be interpreted as a mode effect among first generation immigrants.

Hardcastle and colleagues (2017) found that students for whom English is an additional language (EAL) performed worse on screen than on paper. Wagner and colleagues (2021) found that students who did not mainly speak German in their everyday life performed worse in OSA than in PBA for booklet 1 in the VERA assessments discussed above, but that this performance difference was not statistically significant. In booklet 2, the number of students for whom German was their non-dominant language was too small to make reliable inferences from.

Gallagher and colleagues (2000) did not find any consistent mode effect based on language, but did find that examinees from African American and Hispanic backgrounds performed equivalently or better in OSAs when compared to PBAs. Additionally, using a subset of PISA 2012 mathematics data, Jerrim (2016) investigated whether performance in OSA was influenced by country of birth. He found that the performance gap between first generation immigrants and country natives was significantly different between modes in 12 economies. In 8 instances, this gap between first generation immigrants and country natives was significantly larger in OSAs than in PBAs, although Jerrim notes that the pattern and strength of the relationship was not as consistent or as large as the effect of gender or socio-economic status. It should also be noted that the relationship between ethnicity and immigrant status is not defined, and that Jerrim himself does not outline how, or why, immigrant status may affect performance. Further, socio-economic status and immigrant status were not simultaneously modelled. As such, given that cross-cultural evidence exists to suggest that poverty can be both a cause and consequence of immigration (for example, Aragona and colleagues, 2013; Battistella, 2001; Kesler, 2014; Thiede and colleagues, 2021), it is possible that the effect observed among first-generation immigrants is not independent and separate from socio-economic status. This is the conclusion reached by Wagner and colleagues (2021) in their study of mode effects in the German VERA assessment: they found that students who did not mainly speak German in their everyday life generally performed worse, and emphasise that these students often came from immigrant families and that these families often live in lower socioeconomic contexts.

Socioeconomic status

Some studies have considered if, and how, socioeconomic status (SES) may be linked to mode effects. It is worth stressing from the outset that SES is defined inconsistently which can make comparisons difficult, but – in general terms – each definition relies upon some form of quantifiable economic measure (such as eligibility for free or reduced cost school meals (for example, Horkay and colleagues, 2006)), meaning it can be said that the findings describe mode effects in relation to at least some facets of economic (dis)advantage.

In an experiment to explore mode effects in a test of computing skills with a sample of volunteers from New South Wales aged 15 and 16, MacCann (2005) found a statistically significant interaction between SES and delivery mode. Specifically, he found that high SES students performed comparably in PBA and OSA versions of the same test, whereas the OSA mean mark was 1% lower than the PBA mean mark for low SES students. In their report on mode effects in NAEP, Horkay and colleagues (2006) used repeated measures ANOVA to examine the effect of demographic variables on essay scores. They found that, when those who sat the OSA were compared with those who sat the PBA, their measure related to SES had no statistically significant direct effect, or interacting effect with delivery mode, on performance.

Computer familiarity is closely linked to socioeconomic status and could potentially exacerbate performance gaps between different socioeconomic groups (ITC & ATP, 2022). That being said, it should also be noted that computer familiarity is not necessarily a fixed variable and is context dependent. For example, performance in OSAs can be impacted by taking exams on devices which are unfamiliar, such as those owned by an educational institution (Lee, 2002; Walker & Handley, 2016). Further, research by Harris and colleagues (2017) using a sample of 1,351 Western Australian children suggests that the use of ICT in both the school and the home may differ along socioeconomic lines. Specifically, they found that, compared with participants from lower SES neighbourhoods, participants from higher SES neighbourhoods were more likely to use computers in school and to use computer learning programs, and less likely to use computers at home and for recreational activities. This could suggest that computer familiarity may mean different things to different young people, depending on their socioeconomic context.

Nevertheless, there is mixed evidence as to whether computer familiarity moderates performance in OSA. Russell and Haney (1997) undertook a study of KS3-aged students enrolled in a school where ICT was uniquely integrated into teaching and learning, and they concluded that computer familiarity explained better performance in OSA for extended written responses. However, Leeson (2006) concluded, based upon her review of the literature, that there is inconclusive evidence about the impact of computer familiarity on performance, which is consistent with the findings of this literature review: some studies suggest that computer familiarity impacts performance in national surveys (Bennett and colleagues, 2008; Horkay and colleagues, 2006), summative (Walker & Handley, 2016; Wallace & Clariana, 2005), and one-off experimental assessments (Russell, 1999) across a range educational levels. Other studies found no effect (Clariana & Wallace, 2002; Eid, 2005). The inconsistency of the results was underscored by a meta-analysis conducted by Scherer & Siddiq (2019) using 32 independent K-12 student samples that provided 75 correlation coefficients. They identified a positive, significant, small correlation between SES and ICT literacy, but note that the correlation was moderated by the type of SES measure, the type of ICT literacy assessment, the broad category of ICT skills being assessed, the assessment of test fairness, and the sampling procedure employed. However, if computer familiarity does affect performance, it can have significant implications for assessment fairness.

Underlying ability

The effect of the mode of assessment on students of different ability groups is unclear. Some studies have indicated that OSAs may reduce the attainment gap between the highest- and lowest-achieving students, but there is no consensus on the mechanism through which this occurs. For example, in the context of a formative assessment developed for use in a field experiment, Nikou and Economides (2016) found that low-achieving students performed better on OSA. Relatedly, Ponce and colleagues (2021) undertook experimental research with students at primary and middle-school levels comparing different forms of OSAs: specifically, they compared OSAs where the items were direct copies of the PBA items with OSAs that included technology enhanced items in the form of drag-and-drop items. From their results, Ponce and colleagues suggest that OSA may help low-achieving students perform better, with the authors suggesting that features such as drag-and-drop items can reduce construct irrelevant variance in the form of cognitive demands imposed by conventional computer interfaces.

Other studies, however, suggested that the gap between the highest- and lowest-achieving students might be reduced because OSAs will have a disproportionately detrimental impact on the highest-achieving students. For example, using a subset of PISA 2012 data, Jerrim (2016) found that the gap between high- and low-performing students decreased when assessment was on screen, but found this reduction was a result of the decline in the performance of higher-attainers. Similarly, using data from a study conducted in 2015 to prepare for the digitalisation of the Norwegian National Reading Test, Støle and colleagues (2020) found that, among 10-year-olds, the decline in comprehension with OSA was most pronounced among high-performing females.

In contrast, there is evidence from a writing assessment in primary education (White and colleagues, 2015) and an assessment of general computer-related facts and concepts in a higher education course of computer fundamentals in business (Clariana & Wallace, 2002) that indicates high-performing students perform better in OSAs than in PBAs. Additionally, Walgermo and colleagues (2013, cited in Støle, 2018) found results that suggest screen reading may negatively impact the performance of the lowest-performing students the most. Wagner and colleagues (2021) also found that students with lower academic achievement were more disadvantaged by OSA than those with higher academic achievement. These studies, covering a range of subjects, would suggest that OSA may increase the attainment gap between the most- and least-able students.

Given contradictory evidence on how students of different abilities perform in OSAs, it is perhaps unsurprising that Jerrim and colleagues (2018) found no consistent mode effect between high- and low-performing students in reading, science, or mathematics in PISA 2015 field trial data from Germany, Sweden, and Ireland.

Beyond performance, Karay and colleagues (2015) conducted a study with 266 German medical students and found that underlying ability interacted with assessment mode, with low-performing students guessing significantly more in the OSA than low performing-students sitting a PBA version of the assessment. However, the reasons for this are unclear, meaning further investigation is needed to understand if, and how, this finding is generalisable to other assessment contexts.

Assessment characteristics

Item characteristics

Some evidence exists to suggest that the certain types of items are more likely to elicit mode effects. These mode effects generally occur because of the affordances of the medium, or because of the inherent properties of certain response formats, and their implications for assessment design are considered in detail in a further review (Ofqual, 2025, Making sense of mode effects).

In analysis of the comparability of the paper and computer-based versions of the PARCC assessment, Liu and colleagues (2016) found that, in mathematics, the items most susceptible to mode effects were short constructed response and “fill-in-the-blank” items, although the authors do not make clear on which mode these items were easier. Research undertaken as part of the NAEP technology-based assessment project found that all items were generally more difficult on-screen, but this was particularly the case for constructed response questions when they were compared with multiple choice questions (Bennett and colleagues, 2008; Sandene and colleagues, 2005). For constructed response questions, some studies have found no significant performance differences between OSA and PBA (for example, Sandene and colleagues, 2005[9]; Horkay and colleagues, 2006), or that performance is better on paper (for example, Chen and colleagues, 2011). This is in tension with

[9] Sandene and colleagues report the results of the mathematics and writing tests from the NAEP technology-based assessment project. They found that students appeared to perform worse in the OSA for mathematics, and that – on average – 5% more students responded to a given item correctly on paper than on computer. In the writing test, they found no significant difference in the scores of the extended constructed response item. The role of subject is considered in a later section.

the finding that typed responses are often longer and contain more relevant information, which is discussed in a later subsection. In the context of an undergraduate psychology assessment, Hewson and colleagues (2007) found no difference between modes in a multiple-choice question test. Indeed, when summarising the literature, Wheadon (2007) notes that non-speeded tests that require short answers or objective responses were not liable to substantial mode effects.

Threlfall and colleagues (2007) undertook a detailed consideration of item characteristics and identified that "exploratory trials" are a unique affordance of the digital medium. For example, if an item requires a student to identify a line of symmetry, on paper the student must first identify the line mentally before committing it to paper. Repeated attempts at drawing and erasing the line would create mess and make it harder to answer the question on paper. In contrast, the digital medium might allow students to place and move the line until the correct solution is found. In short, the paper-based version of the item requires the student to know the answer in advance, whereas the digital version does not. In items that allowed "exploratory trials" in responses, Threlfall and colleagues found that performance was better on screen. Contrastingly, they also identified that certain tasks – such as drawing a triangle – were more difficult on screen because of software limitations.

Affordances of the medium become salient when an item that is also delivered on paper is adapted for on-screen delivery and there is a difference in the affordances across mediums. For example, the potential for mode effects in items that require working out has been highlighted, or empirically identified, in a number of instances (Hughes and colleagues, 2010; Johnson & Green, 2006; Keng and colleagues, 2008; Liu and colleagues, 2016; Threlfall and colleagues, 2007).This effect is often attributed to the physical distance between the screen and scratch paper that is generally given to students on which they can perform working out, with it hypothesised that the distance either impacts the willingness of the students to transcribe their working or leads to transcription errors. Keng and colleagues (2008) found that students performed worse on screen in items that involved scrolling when compared with the same items that did not require scrolling in the paper administration. In a reading test involving 1,163 seventh grade students from German schools, Buerger and colleagues (2019) found no difference between modes for multiple choice items, but did find that items which required the use of the use of drop-down menus on screen were more difficult than equivalent items on paper. Wu and colleagues (2015) examined a subset of technology enhanced items: specifically, how performance differed when items were presented dynamically and statically. With a sample of 561 eighth grade students and 657 11th grade students, they found that when items were presented statically, they were easier for the eighth grade students, but more difficult for the 11th grade students. Wu and colleagues hypothesised that this was because older students have more cognitive resources that enable them to navigate dynamic items and assessment tasks simultaneously. This could suggest that the use of technology enhanced items may introduce mode effects, but that these mode effects may differ by age.

Length versus quality of responses

Research suggests that typed responses tend to be longer than handwritten ones (Russel & Haney, 1997; Masterman, 2018; Zehner and colleagues, 2019). In a higher education English test in China, Jin and Yan (2017) found that typed responses were longer, contained longer sentences, and had fewer errors. Analysis of German students' responses to the PISA reading items found that the typed responses in 2015 were slightly longer on average than the handwritten responses in 2012 and contained slightly more relevant information (Zehner and colleagues, 2019). However, Zehner and colleagues (2019) also made several findings which suggest that there is not a linear relationship between typing and quality of responses: they found there was a wide spread in the relative average percentage of relevant information included in PISA 2015 typed responses, ranging from 7% less relevant to 18% more relevant information compared with written responses. In addition, the proportion of relevant information in responses at the item level varied based on the assessment cycle (2012 or 2015) and gender.

As such, while typed essays may often be longer, the length of typed responses does not always reflect students' levels of performance, and is not always a proxy for the quality of responses. Masterman (2018) summarises the literature relating to this point, noting that: there is generally a correlation between longer scripts and higher marks; typed scripts often appear shorter than handwritten scripts of similar or greater length; markers often have higher expectations of typed scripts; but there is no evidence in the literature of clear, systematic difference in the marks awarded to typed scripts. As such, it is difficult to say if assessment mode affects how an item is marked. Notably, Jin and Yan (2017) found that higher scores on typed essays were a result of greater computer familiarity, and Lynch (2022) suggests that Chen and colleagues' (2011) finding of better performance on paper may be explained by the fact that their sample was of older adults who learned to write with a pen and paper. As such, the relationship between performance and the length of typed and handwritten responses may be moderated by other factors such as age and computer familiarity.

Subject

There is evidence that differences in performance in assessments on screen and on paper varies by subject of the assessment. It may be prudent not to think of these subject effects as independent from the effects discussed above – it may be the case that they are product of mode effects caused by screen reading and item-type differences. A random effects meta-analysis, conducted by Kingston (2009), of studies that included elementary, middle, and high school students found that while the mode effect did not vary by school level, the mode effect did vary by subject, with students performing slightly better on an OSA in English (mean effect size of .11) and Social Studies (mean effect size of .15) and slightly better on PBA in maths (mean effect size of -.06). As mentioned, some studies have highlighted issues with transcription of working out as being linked to poorer performance in maths (Hughes and colleagues, 2010; Johnson & Green, 2006; Keng and colleagues, 2008). Some other studies have raised questions about whether differences between the functionality of screen media and pen and paper means that the same item may ultimately test different things in OSAs and PBAs. For example, questions can be raised about the comparability of constructs when figures need to be transcribed on jotting paper so that they can be annotated like those in an exam booklet are, meaning that the OSA items involve an additional step and may reward drawing ability (Keng and colleagues, 2008). Similarly, the techniques identified by Threlfall and colleagues (2007) to reduce cognitive load in maths may arguably impact comparability of constructs.

The mode effect does not only vary between subjects, but evidence exists to suggest it also varies between content areas within subjects. In particular, there is a lot of evidence that this occurs within mathematics. For instance, Logan’s (2015) study involving data from over 800 year 6 Singaporean students completing mathematics items on iPad screens or on paper alongside a paper-based visuospatial instrument suggests that students performed slightly better on the paper version of items covering whole number, algebraic patterns, and data and chance content areas. The better performance was explained by the students’ greater familiarity and ease of showing work and working out solutions on paper than on a screen, especially on a tablet screen. Items where students may need to draw on a graph have been identified as potentially problematic by Liu and colleagues (2016) and Keng and colleagues (2008). In contrast, Boote and colleagues (2021) examined the mode effects in a test of graph comprehension among MBA students and found no difference at the test level, but that students performed better on screen in scatterplot items. Gu and colleagues (2006) undertook a detailed analysis of content areas among graduate students sitting a GRE maths exam. 75% of the items functioned differently across modes, with arithmetic items generally being more difficult on screen and equalities/inequalities and variables items generally being more difficult on paper. Keng and colleagues also found mathematics items requiring geometric manipulations were susceptible to mode effects.

Summary of evidence based on pre-18 Education

As outlined in the introduction, much of the research into OSA has been undertaken at higher education level through experimental methods. Given Ofqual's regulatory activities are focused on the pre-higher education levels, it is worthwhile examining the findings based only on sources which undertook research at the equivalents of KS1, KS2, KS3, and KS4. The picture is similar to the one painted in earlier sections of this review: there are inconsistent patterns of mode effects.

For example, at the primary level, White and colleagues (2015) found that high-performing students perform better on screen, while Jerrim's research provided mixed evidence relating to the performance of 15-year-olds on screen. Jerrim (2016) found that the performance of the most able students in some jurisdictions (notably Shanghai and Chinese Taipei) was worse on screen in PISA 2012 which resulted in a narrower attainment gap between high and low performing students. However, he noted that this effect was not universal, with other countries (notably the USA) showing minimal difference. A similar analysis Jerrim carried out with colleagues on PISA 2015 field trial data again suggested there were no consistent mode effects between the most- and least-able students (Jerrim and colleagues, 2018). Wheadon (2007) suggests his own findings indicate that middling-ability students performed best in OSA because the OSA versions of the biology, physics, and chemistry GCSEs he studied were less speeded than the PBA versions. He bases his argument on the finding that items towards the end of the OSA exhibited non-uniform DIF. However, Keng and colleagues (2008) found no evidence of an item position effect in maths or English. Most studies considered mode effects at test level – when Hughes and colleagues (2007) did this, they found minimal difference between modes, but when they considered differences at item level, they found some significant differences between modes favouring performance in PBA.

Among KS3-aged school students, Russell (1999) found that on-screen performance was better in science, but worse in maths, although the impact in maths was lessened for students with greater keyboarding skill. In English, there was no overall effect, but students with greater keyboarding skill performed moderately better in the OSA. Bennett and colleagues (2008) and Horkay and colleagues (2006) also found that computer familiarity affected performance with children of KS3 age, while at primary level, Eid (2005) found no impact of computer familiarity. King and colleagues (2011) found that screen size had no impact on performance in their study of live testing whereas Bridgeman and colleagues (2003) found that it did in their experimental study with high school juniors. Bridgeman and colleagues (2003) did not find that scrolling affected performance, but Keng and colleagues (2011) did.

Russell and Haney (1997) found the on-screen written responses of KS3-aged students were quantitatively longer, but not necessarily qualitatively better, while Zehner and colleagues (2019) found the longer responses of KS4-aged students also contained more relevant information.

The studies most easily compared with one another – in that they used similar data with samples at the same education level and are from similar points in time – pertain to reading comprehension. But, even there, there are conflicting findings. At KS4 level, some studies suggest female students have poorer comprehension on screen (Støle and colleagues, 2020) and that the gender gap in reading literacy can decrease with screen reading (Borgonovi, 2016). In contrast, Jensen (2020) finds that KS4-aged school females have higher comprehension scores on screen whereas males have higher scores on paper. Jensen also found no difference in comprehension between screen and paper reading, whereas Eyre and colleagues (2017), Støle and colleagues (2018), and Mangen and colleagues (2013) found that comprehension scores are higher on paper.

Conclusion

This review is part of a broader programme of work exploring the opportunities, challenges, and risks of a potential increase in the use of OSA in high-stakes qualifications in England. Specifically, it sought to identify how students' levels of engagement with, cognitive demands experienced during, and performance in assessment may be impacted by a move from PBA to OSA, and if these impacts may unfairly disadvantage specific groups.

This review found some evidence to suggest that students often prefer OSAs. This particularly appears to be the case when students are more familiar with computing and typing. However, this needs to be balanced against findings that indicate some students still prefer handwriting in some instances. Further, other findings point to the possibility that preference with OSA is associated with its novelty and suggest that students may sometimes have negative experiences with OSA, such as increased stress and fatigue, or experiences that arise from the novel distractions associated with OSAs.

Much of the evidence suggests that screen media may pose more cognitive demands than paper-based media, particularly when it comes to reading. This can be the result of various factors, such as the need to scroll or the design of the OSA more broadly. However, there was also evidence to suggest that some of these demands can be mitigated through appropriate assessment design measures, and that OSAs may actually reduce the cognitive demands associated with assessment tasks in certain instances.

The effects of OSA on engagement and cognitive demands only matters insofar as there is an impact on performance in assessment. Here, there is mixed evidence of mode effects. This holds true even when subgroups of students or specific features of the assessment are focused on.

The literature that exists has several limitations. Most significantly, many of the studies undertaken are at the higher education level in the US and have used experimental methods which are typically low stakes for students. The applicability of their findings to high stakes, sessional examinations in England for students typically aged 16 to 18 is unclear. However, even when only evidence from pre-18 education is considered, the evidence of mode effects remained mixed.

As such, caution is needed when trying to apply the findings of the literature to the context of English high stakes exams. A move to OSA likely will be associated with mode effects because of the different affordances associated with this medium. It is not clear, though, what these mode effects will look like, or if they will have a uniform impact across students and different assessments. The evidence suggests that, while some mode effects likely arise as a result of the intrinsic affordances of digital media, the interaction of different factors in the assessment context – including student characteristics, characteristics of the assessment, and the broader social and national context – appear to drive mode effects, too. It is therefore important to take a case-by-case approach, both when assessing the relevance of previous findings, and when assessing the possible effects of any moves to OSA in England, particularly if some students will take OSAs while others take PBAs.

Reference list

Ackerman, R., & Goldsmith, M. (2011). Metacognitive regulation of text learning: on screen versus on paper. *Journal of experimental psychology. Applied*, *17*(1), 18–32. https://doi.org/10.1037/a0022086

Akbay, T., Akbay, L., Erol, O. (2021). Test delivery medium matters: cognitive effort exertion on assessment. *Malaysian Online Journal of Educational Technology, 9*(2), 76–85 https://doi.org/10.52380/mojet.2021.9.2.273

Albuquerque, J., Bittencourt, I. I., Coelho, J. A., & Silva, A. P. (2017). Does gender stereotype threat in gamified educational environments cause anxiety? An experimental study. *Computers & Education*, *115*, 161–170. https://doi.org/10.1016/j.compedu.2017.08.005

Aragona, M., Pucci, D., Mazzetti, M., Maisano, B., and Geraci, S. (2013). Traumatic events, post-migration living difficulties and post-traumatic symptoms in first generation immigrants: a primary care study. *Annali dell'Istituto Superiore di Sanita*, 49, 169–175. doi: 10.4415/ANN_13_02_08

Backes, B., & Cowan, J. (2024). Are Online and Paper Tests Comparable? Evidence from Statewide K-12 Tests. *Applied Measurement in Education*, *37*(1), 1–13. https://doi.org/10.1080/08957347.2024.2311933

Battistella, G. (2001). *International Migration in Asia*. Scalabrini

Benedetto S, Drai-Zerbib V, Pedrotti M, Tissier G, & Baccino, T (2013) E-Readers and Visual Fatigue. PLoS ONE 8(12): e83676. https://doi.org/10.1371/journal.pone.0083676

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it Matter if I Take My Mathematics Test on Computer? A Second Empirical Study of Mode Effects in NAEP. *The Journal of Technology, Learning, and Assessment*, *6*(9). www.jtla.org

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on Computer-Based Test performance. *Applied Measurement in Education*, *16*(3), 191–205. https://doi.org/10.1207/s15324818ame1603_2

Bryant, W. P. (2017). Developing a strategy for using Technology-Enhanced items in Large-Scale standardized tests. *Practical Assessment, Research and Evaluation*, *22*(1), 1. https://doi.org/10.7275/70yb-dj34

Boevé, A. J., Meijer, R. R., Albers, C. J., Beetsma, Y., & Bosker, R. J. (2015). Introducing Computer-Based Testing in High-Stakes Exams in Higher Education:

Results of a Field Experiment. *PloS one*, *10*(12), e0143616. https://doi.org/10.1371/journal.pone.0143616

Boote, S. K., Boote, D. N., & Williamson, S. (2021). Assessing Graph Comprehension on Paper and Computer with MBA Students: A Crossover Experimental Study. *Cogent Education*, *8*(1). https://doi.org/10.1080/2331186x.2021.1960247

Borgonovi F. (2016). Video gaming and gender differences in digital and printed reading performance among 15-year-olds students in 26 countries. *Journal of adolescence*, *48*, 45–61. https://doi.org/10.1016/j.adolescence.2016.01.004

Buerger, S., Kroehne, U., Koehler, C., & Goldhammer, F. (2019). What makes the difference? The impact of item properties on mode effects in reading assessments. *Studies in Educational Evaluation*, *62*, 1–9. https://doi.org/10.1016/j.stueduc.2019.04.005

Cassady, J. C., Budenz-Anders, J., Pavlechko, G., & Mock, W. (2001, December 4). The Effects of Internet Based Formative and Summative Assessment on Test Anxiety, Perceptions of Threat, and Achievement. *Paper Presented at the Annual Meeting of the American Educational Research Association*.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology*, *27*(2), 270–295. https://doi.org/10.1006/ceps.2001.1094

Ceka, B., & O'Geen, A. (2019). Evaluating Student Performance on Computer-Based versus Handwritten Exams: Evidence from a Field Experiment in the Classroom. *PS: Political Science & Politics, 52*(4), 757-762. doi:10.1017/S104909651900091X

Chan, C. K. Y. (2023). A systematic review – handwritten examinations are becoming outdated, is it time to change to typed examinations in our assessment policy? *Assessment & Evaluation in Higher Education*, 1–17. https://doi.org/10.1080/02602938.2023.2219422

Charman, M. (2013). Linguistic analysis of extended examination answers: Differences between on-screen and paper-based, high- and low-scoring answers. *British Journal of Educational Technology*, *45*(5), 834–843. https://doi.org/10.1111/bjet.12100

Chen, J., White, S., McCloskey, M., Soroui, J., & Chun, Y. (2011). Effects of computer versus paper administration of an adult functional writing assessment. *Assessing Writing*, *16*(1), 49–71. https://doi.org/10.1016/j.asw.2010.11.001

Clariana, R. B., & Wallace, P. (2002). Paper–based versus computer–based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology*, *33*(5), 593–602. https://doi.org/10.1111/1467-8535.00294

Clinton, V. (2019). Reading from paper compared to screens: A systematic review and meta-analysis. *Journal of Research in Reading*, *42*(2), 288–325. https://doi.org/10.1111/1467-9817.12269

Connelly, V., Gee, D., & Walsh, E. (2007). A comparison of keyboarded and handwritten compositions and the relationship with transcription speed. *British Journal of Educational Psychology, 77*(2), 479–492. https://doi.org/10.1348/000709906X116768

Dahan Golan, D., Barzillai, M., & Katzir, T. (2018). The effect of presentation mode on children's reading preferences, performance, and self-evaluations. *Computers & Education*, *126*, 346–358. https://doi.org/10.1016/j.compedu.2018.08.001

Delgado, P., Vargas, C., Ackerman, R., & Salmerón, L. (2018). Don't throw away your printed books: A meta-analysis on the effects of reading media on reading comprehension. *Educational Research Review*, *25*, 23–38. https://doi.org/10.1016/j.edurev.2018.09.003

Dillon, A. (1992). Reading from paper versus screens: a critical review of the empirical literature. *Ergonomics*, *35*(10), 1297–1326. https://doi.org/10.1080/00140139208967394

Dutke, S., & Stöber, J. (2001). Test anxiety, working memory, and cognitive performance: Supportive effects of sequential demands. *Cognition & Emotion*, *15*(3), 381–389. https://doi.org/10.1080/0269993004200231

Eid, G. K. (2005). An Investigation into the Effects and Factors Influencing Computer-Based Online Math Problem-Solving in Primary Schools. *Journal of Educational Technology Systems*, *33*(3), 223–240. https://doi.org/10.2190/J3Q5-BAA5-2L62-AEY3

Engelbrecht, J., & Harding, A. (2004). Combing Online and Paper Assessment in a Web-based Course in Undergraduate Mathematics. *Journal of Computers in Mathematics and Science Teaching, 23*(3), 217–231.

Eyre, J., Berg, M., Mazengarb, J., & Lawes, E. (2017). *Mode Equivalency in PAT: Reading Comprehension*. New Zealand Council for Educational Research

Farinosi, M., Lim, C., & Roll, J. (2015). Book or screen, pen or keyboard? A cross-cultural sociological analysis of writing and reading habits basing on Germany, Italy and the UK. *Telematics and Informatics*, *33*(2), 410–421. https://doi.org/10.1016/j.tele.2015.09.006

Fluck, A., Pullen, D., & Harper, C. (2009). Case study of a computer based examination system. *Australasian Journal of Educational Technology*, *25*(4). https://doi.org/10.14742/ajet.1126

Fortunati, L., & Vincent, J. (2013). Sociological insights on the comparison of writing/reading on paper with writing/reading digitally. *Telematics and Informatics*, *31*(1), 39–51. https://doi.org/10.1016/j.tele.2013.02.005

Franic, D. M., Palmer, R., & Fulford, M. (2021). Doctor of pharmacy student preferences for computer-based vs. paper-and-pencil testing in a, social and administrative pharmacy course. *Currents in pharmacy teaching & learning*, *13*(7), 819–825. https://doi.org/10.1016/j.cptl.2021.03.018

Gallagher, A., Bridgeman, B., & Cahalan, C. (2000). The Effect of Computer-Based Tests on Racial/Ethnic, Gender, and Language Groups. In *ETS Research Report Series* (Vol. 2000, Issue 1). Wiley. https://doi.org/10.1002/j.2333-8504.2000.tb01831.x

Garland, K., & Noyes, J. (2004). CRT monitors: Do they interfere with learning? *Behaviour & Information Technology*, *23*(1), 43–52. https://doi.org/10.1080/01449290310001638504

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*(2). www.jtla.org

Halamish, V., & Elbaz, E. (2020). Children's reading comprehension and metacomprehension on screen versus on paper. *Computers & Education*, *145*, 103737. https://doi.org/10.1016/j.compedu.2019.103737

Hardcastle, J., Herrmann-Abell, C. F., & DeBoer, G. E. (2017). Comparing Student Performance on Paper-and-Pencil and Computer-Based-Tests. *Paper Presented at the 2017 AERA Annual Meeting*.

Harris, C., Straker, L., & Pollock, C. (2017). A socioeconomic related "digital divide" exists in how, not if, young people use computers. *PLoS ONE*, 12(3). https://doi.org/10.1371/journal.pone.0175011

He, Q. & El Masri, Y. H. (2025). An exploration of the effect of speededness in a selection of GCSE examinations. (Ofqual research report 25/7267/3). Ofqual. Retrieved from https://www.gov.uk/government/publications/an-exploration-of-the-effect-of-speededness-in-a-selection-of-gcse-examinations

Helsper, E. J., & Eynon, R. (2009). Digital natives: Where is the evidence? *British Educational Research Journal*, *36*(3), 503–520. https://doi.org/10.1080/01411920902989227

Hewson, C., Charlton, J., & Brosnan, M. (2007). Comparing online and offline administration of multiple choice question assessments to psychology undergraduates: Do assessment modality or computer attitudes influence performance? *Psychology Learning & Teaching*, *6*(1), 37–46. https://doi.org/10.2304/plat.2007.6.1.37

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it Matter if I Take My Writing Test on Computer? An Empirical Study of Mode Effects in NAEP. *The Journal of Technology, Learning, and Assessment*, *5*(2). www.jtla.org

Hughes, S., Custodio, I., Sweiry, E., & Clesham, R. (2010). *Beyond multiple choice: Do e-assessment and mathematics add up?*. Edexcel.

International Test Commission and Association of Test Publishers. (2022). *Guidelines for Technology-Based Assessment International Test Commission and Association of Test Publishers*. https://www.intestcom.org/page/28

Jensen, R. E. (2020). Implications of Changing the Delivery Mode on Reading Tests in Norway—A Gender Perspective. In T. S. Frønes, A. Pettersen, J. Radišić, & N. Buchholtz (Eds.), *Equity, Equality, and Diversity in the Nordic Model of Education* (pp. 337–362). Springer.

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. https://doi.org/10.1080/0969594x.2016.1147420

Jerrim, J., Micklewright, J., Heine, J., Sälzer, C., & McKeown, C. R. (2018). PISA 2015: how big is the ‘mode effect’ and what has been done about it? *Oxford Review of Education*, *44*(4), 476–493. https://doi.org/10.1080/03054985.2018.1430025

Jin, Y., & Yan, M. (2017). Computer literacy and the construct validity of a high-stakes computer-based writing assessment. *Language Assessment Quarterly*, 14(2), 101–119. https://doi.org/10.1080/15434303.2016.1261293

Johnson, M., & Green, S. (2006). On-line Mathematics Assessment: The Impact of Mode on Performance and Question Answering Strategies. *The Journal of Technology, Learning and Assessment*, *4*(5). https://ejournals.bc.edu/index.php/jtla/article/view/1652

Karay, Y., Schauber, S. K., Stosch, C., & Schüttpelz-Brauns, K. (2015). Computer versus Paper—Does it make any difference in test performance? *Teaching and Learning in Medicine*, *27*(1), 57–62. https://doi.org/10.1080/10401334.2014.979175

Keng, L., McClarty, K. L., & Davis, L. L. (2008). Item-level comparative analysis of online and paper administrations of the Texas assessment of knowledge and skills. *Applied Measurement in Education, 21*(3), 207-226. https://doi.org/10.1080/08957340802161774.

Kesler, C. (2014). Welfare states and immigrant poverty. *Acta Sociologica*, *58*(1), 39–61. https://doi.org/10.1177/0001699314560238

King, L., Kong, X. J., & Bleil, B. (2011). Does Size Matter? A Study on the Use of Netbooks in K-12 Assessments. *Paper Presented at the Annual Meeting of the American Educational Research Association.*

http://images.pearsonassessments.com/images/PDF/AERA-Netbooks_K-12_Assessments.pdf

Kingston, N. M. (2008). Comparability of Computer- and Paper-Administered Multiple-Choice Tests for K–12 populations: a synthesis. *Applied Measurement in Education*, *22*(1), 22–37. https://doi.org/10.1080/08957340802558326

Kong, Y., Seo, Y. S., & Zhai, L. (2018). Comparison of reading performance on screen and on paper: A meta-analysis. *Computers & Education*, *123*, 138–149. https://doi.org/10.1016/j.compedu.2018.05.005

Langer, D., Weintraub, N., & Erez, A. B. H. (2016). Participation in Activities and Relationship to Musculoskeletal Pain in Elementary School Children. *The Israeli Journal of Occupation Therapy*, 25(3), E63–E80.

Lee, Y. (2002). A comparison of composing processes and written products in timed-essay tests across paper-and-pencil and computer modes. *Assessing Writing*, *8*(2), 135–157. https://doi.org/10.1016/s1075-2935(03)00003-5

Leeson, H. V. (2006) The Mode Effect: A Literature Review of Human and Technological Issues in Computerized Testing. *International Journal of Testing*, *6*(1), 1-24 https://doi.org/10.1207/s15327574ijt0601_1

Lim, E. C., Ong, B. K., Wilder-Smith, E. P., & Seet, R. C. (2006). Computer-based versus pen-and-paper testing: students' perception. *Annals of the Academy of Medicine, Singapore*, *35*(9), 599–603.

Liu, J., Brown, T., Chen, J., Ali, U., Hou, L., & Costanzo, K. (2016). Mode Comparability Study based on Spring 2015 Operational Test Data. https://files.eric.ed.gov/fulltext/ED599049.pdf

Logan, T. (2015). The influence of test mode and visuospatial ability on mathematics assessment performance. *Mathematics Education Research Journal, 27*(4), 423-441. https://doi.org/10.1007/s13394-015-0143-1

Lynch, S. (2022). Adapting Paper-Based Tests for Computer Administration: Lessons Learned from 30 Years of Mode Effects Studies in Education. *Practical Assessment, Research, and Evaluation, 27,* Article 22.

MacCann, R. (2005). The equivalence of online and traditional testing for different subpopulations and item types. *British Journal of Educational Technology*, *37*(1), 79–91. https://doi.org/10.1111/j.1467-8535.2005.00524.x

Makransky, G., Terkildsen, T., & Mayer, R. E. (2019). Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learning and Instruction, 60,* 225–263. https://doi.org/10.1016/j.learninstruc.2017.12.007

Mangen, A. (2008). Hypertext fiction reading: haptics and immersion. *Journal of Research in Reading, 31*(4), 404–419. https://doi.org/10.1111/j.1467-9817.2008.00380.x

Mangen, A. (2010). Lesing – på skjerm eller papir; er det så nøye, da? [Reading – on screen or paper; does it really matter?]. Norsklæreren, 3–10, 16–20.

Mangen, A., Walgermo, B. R., & Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research*, *58*, 61–68. https://doi.org/10.1016/j.ijer.2012.12.002

Martin, R., & Binkley, M. (2009). Gender Differences in Cognitive Tests: A Consequence of Gender-dependent Preferences for Specific Information Presentation Formats? In F. Scheuermann & J. Björnsson (Eds.), *The Transition to Computer-based Assessment: New Approaches to Skills Assessment and Implications for Large-scale Testing* (pp. 75–82). Office for Official Publications of the European Communities.

Masterman, E. (2018). Typed versus handwritten essay exams: is there a need to recalibrate the gauges for digital assessment?. In M. Campbell, J. Willems, C. Adachi, D. Blake, I. Doherty, S. Krishnan, S. Macfarlane, L. Ngo, M. O'Donnell, S. Palmer, L. Riddell, I. Story, H. Suri, & J. Tai (Eds.), *ASCILITE 2018* (pp. 204–213). Australasian Society for Computers in Learning in Tertiary Education.

Mavridis, A., & Tsiatsos, T. (2016). Game-based assessment: investigating the impact on test anxiety and exam performance. *Journal of Computer Assisted Learning*, *33*(2), 137–150. https://doi.org/10.1111/jcal.12170

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449–458. https://doi.org/10.1037/0033-2909.114.3.449

Mogey, N., & Hartley, J. (2013). To write or to type? The effects of handwriting and word-processing on the written style of examination essays. *Innovations in Education and Teaching International*, *50*(1), 85–93. https://doi.org/10.1080/14703297.2012.748334

Nikou, S. A., & Economides, A. A. (2016). The impact of paper-based, computer-based and mobile-based self-assessment on students' science motivation and achievement. *Computers in Human Behavior*, *55*, 1241–1248. https://doi.org/10.1016/j.chb.2015.09.025

Norman, D. (2002) Affordances and Design. Online. Retrieved: December 10, 2024, from: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=538961f9812c70d20c1f48042bab1e8060366ca2

Noyes, J., & Garland, K. (2003). VDT versus paper-based text: reply to Mayes, Sims and Koonce. *International Journal of Industrial Ergonomics*, *31*(6), 411–423. https://doi.org/10.1016/s0169-8141(03)00027-1

Ofcom. (2022). Children and Parents: Media Use and Attitudes Report 2022. Online. Accessed May 7 2025, from: Children and parents: media use and attitudes report 2022

Ofcom. (2023). Children and Parents: Media Use and Attitudes Report 2023. Online. Accessed May 7 2025, from: Children and Parents: Media Use and Attitudes 2023

Ofcom. (2024). *Children and Parents: Media Use and Attitudes Report 2024*. Online. Retrieved February 6 2025, from: Childrens Media literacy report 2024

Ofqual. (2020). *Online and OSA in High Stakes, Sessional Qualifications: A Review of the Barriers to Greater Adoption and how these Might be Overcome*. Online. Retrieved October 10, 2024, from: Online and OSA in high stakes, sessional qualifications. (publishing.service.gov.uk)

Ofqual. (2023). *Perceptions of OSAs Survey Results*. Online. Retrieved February 6, 2025, from: Perceptions_of_on-screen_assessments.xlsx

Ofqual. (2025). *On-Screen Assessment Research Study: Opportunities, benefits, risks and challenges of on-screen assessments in England*. Online. Retrieved 11 December, 2025, from: On-screen assessment research study: Opportunities, benefits, risks and challenges of on-screen assessments in England

Ofqual. (2025). *Making Sense of Mode Effects*. Online. Retrieved 11 December, 2025, from: Making sense of mode effects

Ortner, T. M., & Caspers, J. (2011). Consequences of test anxiety on adaptive versus fixed item testing. *European Journal of Psychological Assessment*, *27*(3), 157–163. https://doi.org/10.1027/1015-5759/a000062

Pieschl, S. (2009). Metacognitive calibration—An extended conceptualization and potential applications. *Metacognition and Learning, 4*(1), 3–31. https://doi.org/10.1007/s11409-008-9030-4

Poggio, J., Glasnapp, D. R., Yang, X., Poggio, A. J., Russell, M., Glasnapp, J., Yang, D. R., & Poggio, X. (2005). A Comparative Evaluation of Score Results from Computerized and Paper & Pencil Mathematics Testing in a Large Scale State Assessment Program. *The Journal of Technology, Learning, and Assessment*, *3*. http://www.jtla.org

Ponce, H. R., Mayer, R. E., & Loyola, M. S. (2021). Effects on Test Performance and Efficiency of Technology-Enhanced Items: An Analysis of Drag-and-Drop Response Interactions. *Journal of Educational Computing Research*, *59*(4), 713–739. https://doi.org/10.1177/0735633120969666

Prensky, M. (2001a). Digital Natives, Digital Immigrants: Part 1. *On the Horizon*, *9*(5), 1-6. https://doi.org/10.1108/10748120110424816

Prensky, M. (2001b). Digital Natives, Digital Immigrants Part 2: Do They Really Think Differently?. *On the Horizon*, *9*(6), 1-6. https://doi.org/10.1108/10748120110424843

Ricketts, C., & Wilks, S. (2002). Improving Student Performance Through Computer-based Assessment: Insights from recent research. *Assessment & Evaluation in Higher Education*, *27*(5), 475–479. https://doi.org/10.1080/0260293022000009348

Revuelta, J., Ximénez, M. C., & Olea, J. (2003). Psychometric and psychological effects of item selection and review on computerized testing. *Educational and Psychological Measurement, 63*(5), 791–808. https://doi.org/10.1177/0013164403251282

Russell, M., & Haney, W. (1997). Testing Writing on Computers: An Experiment Comparing Student Performance on Tests Conducted via Computer and via Paper-and-Pencil. *Education Policy Analysis Archives*, *5*(3). http://nis.accel.worc.k12.ma.us

Russell, M.K. (1999). *Testing on computers: A follow-up study comparing performance on computer and on paper* [Doctoral Dissertation], Boston College. https://www.learntechlib.org/p/127332/.

Sanchez, C. A., & Wiley, J. (2009). To scroll or not to scroll: scrolling, working memory capacity, and comprehending complex texts. *Human factors*, *51*(5), 730–738. https://doi.org/10.1177/0018720809352788

Sandene, B., Horkay, N., Bennett, R. E., Allen, N., Braswell, J., Kaplan, B., & Oranje, A. (2005). Online Assessment in Mathematics and Writing: Reports from the NAEP Technology-Based Assessment Project, Research and Development Series. NCES 2005-457. *National Center for Education Statistics*. http://files.eric.ed.gov/fulltext/ED485780.pdf

Scherer, R., & Siddiq, F. (2019). The relation between students' socioeconomic status and ICT literacy: Findings from a meta-analysis. *Computers & Education*, *138*, 13–32. https://doi.org/10.1016/j.compedu.2019.04.011

Shermis, M. D., & Lombard, D. (1998). Effects of computer-based test administrations on test anxiety and performance. *Computers in Human Behavior*, *14*(1), 111–123. https://doi.org/10.1016/s0747-5632(97)00035-6

Selwyn, N. (2009). The Digital Native: Myth and Reality. *Aslib Proceedings*, *61*(4), 364-379. https://doi.org/10.1108/00012530910973776

Singer, L. M. (2014). Reading Across Mediums. *Presentation at a meeting at the Educational Psychology Colloquium*, Department of Human Development and Quantitative Methodology, College Park, MD.170L.

Singer, L. M., & Alexander, P. A. (2017A). Reading across Mediums: Effects of reading digital and print texts on comprehension and calibration. *Journal of Experimental Education*, *85*(1), 155–172. https://doi.org/10.1080/00220973.2016.1143794

Singer, L. M., & Alexander, P. A. (2017B). Reading on Paper and Digitally: What the Past Decades of Empirical Research Reveal. *Review of Educational Research*, *87*(6), 1007–1041. https://doi.org/10.3102/0034654317722961

Singer Trakhman, L. M., Alexander, P. A., & Berkowitz, L. E. (2017). Effects of Processing Time on Comprehension and Calibration in Print and Digital Mediums. *Journal of Experimental Education*, *87*(1), 101–115. https://doi.org/10.1080/00220973.2017.1411877

Støle, H., Mangen, A., Frønes, T., & Thomson, J. (2018). Digitisation of Reading Assessment. In M. Barzillai, J. Thomson, S. Schroeder, & P. van den Broek (Eds.), *Learning to Read in a Digital World* (pp. 205–224). John Benjamins Publishing Company.

Støle, H., Mangen, A., & Schwippert, K. (2020). Assessing children's reading comprehension on paper and screen: A mode-effect study. *Computers & Education*, *151*, 103861. https://doi.org/10.1016/j.compedu.2020.103861

Taylor, A. R. (2005). *A future in the process of arrival: Using computer technologies for the assessment of student learning*. Society for the Advancement of Excellent in Education.

Terzis, V., & Economides, A. A. (2011). The acceptance and use of computer based assessment. *Computers & Education*, *56*(4), 1032–1044. https://doi.org/10.1016/j.compedu.2010.11.017

Thiede, B. C., Brooks, M. M., & Jensen, L. (2021). Unequal from the start? poverty across immigrant generations of Hispanic children. *Demography*. https://doi.org/10.1215/00703370-9519043

Threlfall, J., Pool, P., Homer and Swinnerton, B. (2007). Implicit aspects of paper and pencil mathematics assessment that come to light through the use of the computer. *Educational Studies in Mathematics, 66*(3), 335-348. https://doi.org/10.1007/s10649-006-9078-5

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology / Psychologie canadienne, 26*(1), 1–12. https://doi.org/10.1037/h0080017

Turan, Z., & Meral, E. (2018). Game-Based versus to Non-Game-Based: The Impact of Student Response Systems on Students' Achievements, Engagements and Test Anxieties. *Informatics in Education*, *17*(1), 105–116. https://doi.org/10.15388/infedu.2018.07

Walgermo, B. R., Mangen, A., & Brønnick, K. (2013).
Lesing av sammenhengende tekster på skjerm og papir: Apropos digitalisering av leseprøver. *Paper Presented at Skriv! Les!*.

Wang, S., Hong, J., Young, M. J., Brooks, T., & Olson, J. (2007). A Meta-Analysis of Testing Mode Effects in Grade K-12 Mathematics Tests. *Educational and Psychological Measurement*, *67*(2), 219–238. https://doi.org/10.1177/0013164406288166

Wagner, I., Loesche, P., & Bißantz, S. (2021). Low-stakes performance testing in Germany by the VERA assessment: analysis of the mode effects between computer-based testing and paper-pencil testing. *European Journal of Psychology of Education*, *37*(2), 531–549. https://doi.org/10.1007/s10212-021-00532-6

Walker, R., & Handley, Z. (2016). Designing for learner engagement with computer-based testing. *Research in Learning Technology*, *24*(1), 30083. https://doi.org/10.3402/rlt.v24.30083

Wallace, P., & Clariana, R. B. (2005). Gender Differences in Computer-Administered versus Paper-Based Tests. *International Journal of Instructional Media*, *32*.

Wästlund, E., Reinikka, H., Norlander, T., & Archer, T. (2005). Effects of VDT and paper presentation on consumption and production of information: Psychological and physiological factors. *Computers in Human Behavior, 21*(2), 377–394. https://doi.org/10.1016/j.chb.2004.02.007

Wästlund, E., Norlander, T., & Archer, T. (2008). The effect of page layout on mental workload: A dual-task experiment. *Computers in Human Behavior*, *24*(3), 1229–1245. https://doi.org/10.1016/j.chb.2007.05.001

Wheadon, C. (2007). *The Comparability of Onscreen and Paper and Pencil Tests: No Further Research Required?*. AQA Centre for Education Research and Policy. The comparability of onscreen and paper and pencil tests: no further research required? (aqa.org.uk)

White, S., Yee Kim, Y., Chen, J., & Liu, F. (2015). *Performance of fourth-grade students in the 2012 NAEP computer-based writing pilot assessment: Scores, text length, and use of editing tools*. National Center for Education Statistics.

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. https://doi.org/10.1016/j.asw.2008.03.001

Wise, S. L. (2019). Controlling construct-irrelevant factors through computer-based testing: disengagement, anxiety, & cheating. *Education Inquiry*, *10*(1), 21–33. https://doi.org/10.1080/20004508.2018.1490127

Wu, H., Kuo, C., Jen, T., & Hsu, Y. (2015). What makes an item more difficult? Effects of modality and type of visual information in a computer-based assessment of scientific inquiry abilities. *Computers & Education*, *85*, 35–48. https://doi.org/10.1016/j.compedu.2015.01.007

Zehner, F., Goldhammer, F., Lubaway, E., & Sälzer, C. (2018). Unattended consequences: how text responses alter alongside PISA's mode change from 2012 to 2015. *Education Inquiry*, *10*(1), 34–55. https://doi.org/10.1080/20004508.2018.1518080

Appendix 1

Table 1: Sources cited that make an empirical claim about mode effects based on research undertaken at the equivalent of Key Stage 1 & Key Stage 2, Key Stage 3, Key Stage 4 and further education levels

	KS1 & KS2	KS3	KS4	Further Education	Total
Formative Assessments					0
Summative Assessments			- Wheadon (2007)		1
Formative and Summative Assessments					0
National and International Surveys Assessment Data		- Wagner and colleagues (2021)	- Borgonovi (2016) - Jerrim (2016) - Jerrim and colleagues (2018) - Zehner and colleagues (2019)		5

	KS1 & KS2	**KS3**	**KS4**	**Further Education**	**Total**
National and International Surveys Field Studies, Trials, and Pilots	- White and colleagues (2015)	- Buerger and colleagues (2019) - Bennett and colleagues (2008) - Jensen (2020) - Horkay and colleagues (2006) - Sandene and colleagues (2005) - Støle and colleagues (2020)			7
Experiments	- Connelly and colleagues (2007) - Dahan Golan and colleagues (2018) - Eid (2005) - Halamish & Elbaz (2020) - Johnson & Green (2006)	- Logan (2015) - Ling and colleagues, (2017) - Russell and Haney (1997) - Russell (1999)	- Bridgeman and colleagues (2003) - Hughes and colleagues (2010) - MacCann (2005) - Mangen and colleagues (2013) - Turan & Meral (2018) - Walgermo and colleagues (2013)	- Charman (2013)	16
Field Experiments, Trials, and Pilots		- Poggio and colleagues (2005)	- Nikou & Economides (2016)		2
Total	6	12	12	1	**31**

Table 2: Sources cited that make an empirical claim about mode effects based on research undertaken at higher education, multiple education levels, and none-education and adult education contexts

	Higher Education	Multiple Education Levels	None-Education and Adult Education Contexts	Total
Formative Assessments		- Eyre and colleagues (2017)		1
Summative Assessments	- Jin & Yan (2017) - Lim and colleagues (2006) - Walker & Handley (2016)	- Backes & Cowan (2024) - Gallagher and colleagues (2000) - Keng and colleagues (2008) - King and colleagues (2011)		7
Formative and Summative	- Engelbrecht & Harding (2004) - Ricketts & Wilks (2002)			2
Experiments	- Ackerman & Goldsmith (2011) - Akbay and colleagues (2021) - Boote and colleagues (2021) - Gu and colleagues (2006) - Lee (2002) - Makransky and colleagues (2019) - Mavridis & Tsiatsos (2016) - Mogey & Hartley (2013)	- Threlfall and colleagues (2007)	- Albuquerque and colleagues (2017) - Benedetto and colleagues (2013) - Garland & Noyes (2004) - Hardcastle and colleagues (2017) - Ortner & Caspers (2011) - Ponce and colleagues (2021)	23

	Higher Education	Multiple Education Levels	None-Education and Adult Education Contexts	Total
	- Noyes & Garland (2003) - Revuelta and colleagues (2003) - Sanchez & Wiley (2009) - Singer & Alexander (2017A) - Singer Trakhman and colleagues (2017) - Terzis & Economides (2011) - Wästlund and colleagues (2005) - Wästlund and colleagues (2008)			
Field Experiments and Trials	- Boevé and colleagues (2015) - Cassady and colleagues (2001) - Ceka & O'Green (2019) - Clariana & Wallace (2002) - Fluck and colleagues (2009) - Franic and colleagues (2021) - Hewson and colleagues (2007) - Karay and colleagues (2015) - Shermis & Lombard (1998) - Wallace & Clariana (2005) - Whithaus and colleagues (2008)		- Chen and colleagues (2011) - Liu and colleagues (2016) - Wu and colleagues (2015)	14
Total	31	6	9	**47**

Table 3: List of cited Reviews and Meta-analyses

	Reviews	Meta-Analyses	Reviews and Meta-Analyses
	- Chan (2023) - Kingston (2009) - Leeson (2006) - Lynch (2022) - Mangen (2008) - Masterman (2018) - Singer & Alexander (2017B) - Støle and colleagues (2018) - Taylor (2005) - Wise (2019)	- Delgado and colleagues (2018) - Kong and colleagues (2018) - Mead & Drasgow (1993) - Scherer & Siddiq (2019) - Wang and colleagues (2007)	- Clinton (2019)
Total	10	5	1

Published by:

ofqual

2nd floor, 1 Friargate
Station Square
Coventry
CV1 2GN

0300 303 3344

public.enquiries@ofqual.gov.uk

www.gov.uk/ofqual

December 2025

Ofqual/25/7280