RESEARCH AND ANALYSIS

Time limits and speed of working in assessments

When, and to what extent, should speed of working be part of what is assessed?

Authors

- Stephen Holmes

How to cite this publication

Holmes, S. (2025). *Time limits and speed of working in assessments: When, and to what extent, should speed of working be part of what is assessed?* (Ofqual research report 25/7267/2). Ofqual. Retrieved from https://www.gov.uk/government/publications/time-limits-and-speed-of-working-in-assessments

Contents

Executive summary

Most assessments in England are taken under timed conditions. For exams this is important for scheduling purposes, while in some cases time limits may be used to set reasonable expectations for how long assessment tasks should take. If the speed at which work is done is not part of what a test is designed to assess (known as the test construct) then the time limit should be sufficient for all (or nearly all) test takers to attempt all questions or tasks on the test without undue haste. If the speed at which work is completed becomes a factor, so that test outcomes are a combination of speed of working and accuracy or quality of response, then the test is said to be speeded.

Examinations in general qualifications in England (such as GCSEs, AS and A levels) are not intended to be speeded. They are tests of knowledge, skills and understanding only, and none of the content requirements published by the Department for Education, the conditions and guidance published by Ofqual or specification documents published by the exam boards mention speed as part of what is intended to be assessed. The same is true of a variety of tests of knowledge, skills and understanding in vocational and technical qualifications that are similar in design to GCSEs, AS and A levels. For all such exams and tests, the time limits are set through experience, precedent and some checks by senior examiners that the time allowed is thought to be sufficient. Whether for some of these tests there is an implicit understanding that speed of working is important is unknown. But no detailed experimental studies appear to have been carried out to investigate the effect of time limits on outcomes of these types of tests.

While exams and written tests perhaps tend to have a higher profile with the public, there are a variety of other forms of assessment, for some of which the speed at which work is completed may be a core part of the test construct. Tests in qualifications that certify skills-based occupational competence may use time limits to assure that test takers can complete work-relevant tasks in a time expected of someone who has reached the targeted proficiency level. Other tests, such as those often used in medical licensing that may combine the deployment of knowledge with some more practical skills, may intentionally use quite tight time limits to check that individuals can work under the kind of time pressure typical in the occupation. All of these assessments tap into the idea of fluency, where higher competence may be indicated not only by task accuracy, but also by the speed of working. In most of these kinds of assessment, speed is not rewarded directly, just task completion within the time limit and an evaluation of the outcome quality. Although not common in assessments in England, it is also possible to include the time in which a task is completed as part of the scoring, and so to directly credit speed.

It is important from a fairness perspective to be clear and explicit when speed is intended to be measured as part of a test construct. Test takers need to know how they will be assessed, or how fast they are expected to work. Also, one of the most common test adjustments given to individuals with a disability or other need that affects their speed of working, is extra time to complete the test. For this adjustment to be fair in tests that are not intended to assess speed of working, those not receiving extra time should all be able to compete the test within the standard time, while those individuals who struggle to finish within the standard test time should receive extra time. If some of those without extra time are not able to finish the test, then they may end up disadvantaged relative to those with extra time.

It is important to know whether tests that are not intended to be speeded are in fact speeded. Any measure or estimate of test speededness is a snapshot of that test and that cohort of test takers. It is an analysis of a specific set of test data, not strictly a property of the test. Therefore, to be confident a test is speeded, the cohort needs to be highly representative, as does the test version, assuming the test comes in different forms (such as different maths papers each summer). Alternatively multiple sets of data across different cohorts and test versions should be analysed.

Recognising this limitation, a variety of methods have been developed to measure or estimate speededness. The simplest and most direct measure of the impact of time is to administer a test under different time limits to 2 equivalent groups. However, this is costly, partly because it cannot easily be done in a live testing situation, but has to be fully experimental, and can require substantial groups of test takers. Because this experimental situation is also likely to be a low-stakes situation for the test takers, their motivation to give their best effort on the test may be low. Sometimes sections of live tests that are not used for scoring (they may be used for trialling new items) may lend themselves to experimental manipulations, but that opportunity is dependent on the structure of the test, and is not usual in England.

A wide variety of methods have also been developed to analyse single test administration data. Some are based on analysing not-reached items, which are non-attempted items at the end of a test, assumed not to have been attempted due to time limitations (see, for example, He and El Masri, 2025 and Walland, 2024 who found mixed evidence of speededness in GCSE exams). Other methods have used reduced response accuracy towards the end of a test to classify test takers as speeded. Where tests are administered on-screen and item timing is available, low item response time, indicating guessing on a multiple-choice test, can be used to identify the point in the test at which test takers begin to run out of time. Quite advanced statistical models are now in existence that can use both item response time and item accuracy to model the behaviour of test takers and the functioning of the test.

However, many of the more recent single-administration methods are designed to analyse tests made up of one-mark multiple-choice questions presented on screen, and do not lend themselves to the assessment context in England, where most tests are pen-and-paper tests, and many items require a written response, including extended writing tasks. Some analyses of question papers such as maths and sciences, which have a large number of questions with only a few marks, have been carried out, largely using analyses of not-reached items at the end of the tests. These have shown only a small degree of speededness on average.

While analysis of unreached items on papers with a lot of items may be feasible, the effects of limited time may be quite subtle on any written test, but particularly on extended-response type items. Under time pressure test takers may spend less time planning, hurry their response or generally write less than they might have with more time. Of course, that might not reduce their marks – there may be a temptation to write more than is necessary and include material in an answer that is not rewardable, just in case it may gain marks. All this means that measuring the speededness of essay-based tests is extremely difficult without multiple test administration methods. However, these methods are logistically difficult in the context of GCSEs, AS and A levels and some vocational and technical qualifications in England, where papers are kept under high security and sat only once then retired.

If speed of working is not intended to be part of the test construct, then further work and thinking may be required to ensure that current time limits are not leading to reduced scores for a substantial number of test takers to an unwarranted extent, especially those not receiving an extra time adjustment. If speed of working were decided to be a valuable attribute to measure in some tests, technology (for example eye-tracking, keystroke logging) may allow ways to measure how long it took to read the question, plan a response, then respond. However, pen-and-paper tests do not offer these options. Moreover in such tests handwriting speed itself may be a limiting factor. Writing speed cannot ever be relevant to the test construct except when the aim is to directly assess writing speed. Use of computer-based testing may allow control over item timing, and even control of the time allowed for specific stages of the question-answering process that are judged to be the most appropriate places to test speed of working.

Introduction

Many assessments serving a variety of purposes are administered under timed conditions. This can vary from tests of occupational competency focussing on practical skills to tests of more academic knowledge, skills and understanding, including in general qualifications such as GCSEs, AS and A levels in England. There are often sound logistical and administrative reasons for specifying a fixed test duration. Carrying out tests under standardised conditions is a desirable feature that has traditionally been thought to promote validity – if testing conditions vary across individuals, any comparisons made and conclusions drawn from test outcomes may be undermined.

A question related to test time limits is whether a test should include the speed at which test takers[1] work as part of the test construct (what the test is intended to measure). In some circumstances the speed at which a task can be completed, whether mental or practical, may be important to the purpose of the test. Speed may be factored into outcomes either through those who work faster being able to attempt more of the assessment tasks, or less commonly, through a direct reflection of the time taken to complete a task on scores. In other tests, speed of working will not be important, it will just be the quality of the test performance that matters, and in this case speed is said to be construct-irrelevant. Ideally, the time set for an assessment that is not designed to test speed of working should be sufficient for most (if not all) test takers to answer all of the questions or complete all of the tasks it contains. If insufficient time is allowed, then the speed at which each test taker works becomes a factor that affects their test outcomes, and the test is said to be ‘speeded’. Those who work faster may score more highly, but this is not what the test was specified and designed to do.

GCSEs, AS and A levels in England are clear instances where the exams they use are not explicitly intended to assess speed of working. Certainly, speed of working is not listed in the subject content published by the Department for Education (for example DfE, 2013a, b) or the subject-level conditions and requirements, including the assessment objectives, published by Ofqual (for example, Ofqual, 2015). Accordingly, speed of working in exams is not an expectation of specification documents published by exam boards. However, some test takers may feel time pressure in exams in England. Perceived time pressure does not necessarily indicate that the exams are speeded because the perception may arise simply through the high-stakes nature of these tests and the pressure this puts on test takers. While individuals may feel that more time would have allowed them to obtain

[1] The term ‘test taker’ is used throughout this report to describe individuals sitting an assessment, rather than using a variety of terms specific to a context such as ‘student’, ‘candidate’ or ‘apprentice’.

a higher score, this may not always be true, since in practice they may not have actually improved their scores with more time.

Beyond informal perceptions of time pressure, there is little information available regarding the speededness or not of exams in GCSEs, AS and A levels because historically this has not been a focus of research in this country. The duration of these exams is set based on a variety of sources of information, such as how long equivalent tests have been in the past, rules of thumb such as ‘a mark a minute’ or the experience and expectations of the senior examiners involved in designing the tests.

Ofqual has recently been considering extra time in exams, which is one form of reasonable adjustment in assessments (see Cuff, Keys, Churchward et al., 2025). These adjustments are part of the legal obligation the Equality Act (2010) places on schools and colleges as well as awarding organisations to ensure that disabled students are not disadvantaged when taking an assessment. This work has prompted thinking around the setting of time limits in assessments and the impact they may have on performance, and when speed of working may, or may not, be a relevant part of a test construct.

Helpfully, in 2020 the National Council on Measurement in Education (NCME) published a book (Margolis and Feinberg, 2020[2]) which summarises most of the research and theorising that has taken place regarding time limits in assessments and test speededness. This book, and the research and theories it details, are focussed on the assessment context in the US, since that is where NCME operates. For general qualifications (academic, subject-based qualifications testing knowledge, skills and understanding[3]), the US context is quite different from that in England, with many tests there comprised of multiple-choice items, and a substantial amount of testing taking place on computers. Only some tests (often college entry exams or selection tests) have questions that require constructed-response answers that may be similar to the type of questions commonly used in written exams in GCSEs, AS and A levels. However, the research the NCME book details provides a sound base from which to consider issues in the English context.

The NCME book also provides some insight into the opportunities presented by computer-based testing (CBT) when considering the measurement of test speededness, which may be useful when we look to the future here in England. Although there is currently provision for completing written exams in GCSEs, AS and A levels by typing responses using laptops for those with disabilities or those for

[2] The electronic version of this book is currently available under an open access license – please refer to the reference list for the link.

[3] Henceforth these will generally be referred to as ‘knowledge-based tests’ to distinguish them from more practical vocational and technical ‘skills-based tests’

whom this is their normal way of working, it is unclear if and when these exams may be routinely carried out as full on-screen assessments. Some tests in vocational and technical qualifications are already offered as on-screen assessments, and the use of this testing mode is growing.

When tests are administered using bespoke CBT platforms, detailed data can be collected on the way candidates perform on the individual items, including the time taken and the order in which items are attempted, which is not possible for handwritten tests. This has led to a particular focus on data analysis and use of statistical models to measure and predict the effect of time limits on test performance in the US, particularly those composed of multiple-choice items. Some of these analytical approaches may have wider application in England in the future if CBT is more widely adopted here, in which case the Margolis and Feinberg book will become a useful reference source.

Finally, this report is not limited just to consideration of the impact of time limits on exams in GCSEs, AS and A levels, although these assessments are central to the arguments presented here. Other assessment methods such as those that may be used in vocational and technical qualifications and apprenticeship end-point assessments are also considered, to widen the focus and think about what time limits and speed of working may mean in assessments beyond just written exams.

This report first considers the question of when speed of working may or may not be a valid part of a test, depending on what the test intends to measure. A variety of different tests are considered where speed of working is important, or where it may be construct-irrelevant. The report then moves on to consider written tests of knowledge, skills and understanding such as those used in GCSEs, AS and A levels, with a detailed consideration of what speed of working means in these tests, and whether it should, or could, be part of the test construct. Methods of estimating the speededness of tests are then considered, which have historically focussed on methods applicable to multiple-choice tests, and particularly those completed onscreen more recently. Some consideration is given of measuring speededness in other types of assessments. Much of the research in this area is from the US, which has a very different assessment context. All of this evidence is then used to help consider the assessment context in England, especially that of GCSEs, AS and A levels, and how speededness could be measured, and speed of working integrated into testing, if desired. The report also offers some thoughts on future directions for research.

The question of time: When is speed of working or responding a valid part of an assessment construct?

Most assessments are carried out under timed conditions. Sometimes these time limits may be generous enough that no individual test taker is limited in their ability to perform to their maximum. But in many cases the time limit may force a test taker to work through test items or tasks at a higher pace than they would choose if time was not limited, and may even lead to them omitting some items for time reasons. Across the whole range of assessment types across different subjects, topics or occupations, a decision needs to be made as to whether the speed at which an individual completes tasks is a key part of what the test is measuring – is it construct-relevant?

A good place to start when considering the setting of time limits for tests is The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014). This provides a set of professional guidelines, published by the American Educational Research Association, the American Psychological Association and the National Council on Measurement in Education. These standards provide an important reference for the design, development, delivery and review of valid tests. They are extensive and represent a consensus view (from North America) of all the considerations that should feed into test design and development. Standard 4.14 states:

> “For a test that has a time limit, test development research should examine the degree to which scores include a speed component and should evaluate the appropriateness of that component, given the domain the test is designed to measure.
>
> Comment: At a minimum, test developers should examine the proportion of examinees who complete the entire test, as well as the proportion who fail to respond to (omit) individual test questions. Where speed is a meaningful part of the target construct, the distribution of the number of items answered should be analyzed to check for appropriate variability in the number of items attempted as well as the number of correct responses. When speed is not a meaningful part of the target construct, time limits should be determined so that examinees will have adequate time to demonstrate the targeted knowledge and skill” (page 90)

Although this standard applies most directly to exam-like tests with multiple items, the same considerations can be made around more occupationally-related assessments where, for example, an assessment may involve the completion of only a single task. If this task is not intended to be time limited, how many test takers failed to complete the task should be considered, and whether the time limit affected the quality of work of those who did complete it.

Other frameworks from outside the US, such as the European Framework of Standards for Educational Assessment (AEA-Europe, 2022), say less about time limits or speededness. Under the ‘admin and logistics’ category the European framework states that:

> “There must also be a level playing field so that test takers have the same chance to demonstrate what they know and can do. Any differences in their performances should be related to the construct which is the focus of the assessment.” (page 10)

Additionally, it notes that to achieve a level playing field, there should be some form of evidence for how time allocations were decided, such as a pilot test. This less detailed statement around setting time limits may reflect the more concise nature of the European Framework compared to the North American Standards, rather than any difference in its perceived importance.

The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) defines speededness as, “The extent to which test takers’ scores depend on the rate at which work is performed as well as on the correctness of the responses. The term is not used to describe tests of speed” (page 223). This points towards a dichotomy frequently made in US theorising and research, originally formalised by Gulliksen (1950), between 2 very different types of tests:

- Power tests: that is, tests of knowledge and understanding in which there should be no time limit on giving an answer as this would constrain the demonstration of what a test taker knows and can do. Accuracy of response across all items is the outcome measure[4].
- Speed tests: that is, tests of processing speed, where a large number of (usually) low-difficulty items are presented that all test takers are expected to be able to answer correctly, and the task is to answer as many as possible in the time available. The number of items responded to is the outcome measure.

Under this particular theoretical framework, academic, knowledge-based tests should be administered as power tests. Speed tests are generally psychological in nature rather than tests of educational outcomes or academic achievement. They may measure speed of responding related to psychological processing or

[4] Some power tests use items of progressively increasing difficulty, in which case the difficulty of the last item reached is the outcome measure.

perception, such as visual, auditory or phonological processing, or skills that include motor functioning, such as typing speed.

It is important to recognise that these are theoretical ideals. The reality is that most tests have time limits and so the speed/power distinction is more of a continuum, with speed of responding a factor in performance on tests (even when it is not stated explicitly to be so) to varying extents. The term usually used in the US literature for tests with set durations is ‘time-limit tests’. Their outcomes are based on the accuracy of responses to those items that were responded to within the available time.

Some authors in the US have taken a purist view, that to be a valid test of knowledge and understanding, speed should absolutely not be part of the test construct, and that any speededness in time-limit tests represents a threat to their validity (e.g. Lu and Sireci, 2007). A recent article by Gernsbacher et al (2020) states the strong view that time-limit tests are less valid, reliable, inclusive and equitable than untimed power tests. This is not a new view. Davidson and Carroll (1945) made the following observation regarding what they call psychological tests (which includes educational testing):

> “The indiscriminate use of time-limit scores is one of the most unfortunate characteristics of current psychological testing since the time-limit score of a test frequently represents two relatively independent aspects of behavior:
>
> (a) the amount the subject [sic] knows or can perform (or in certain cases, the level of difficulty which he [sic] can reach) and
>
> (b) the rate at which the subject [sic] works.” (page 411)

However, in the US many tests are taken in paid-for test centres, and longer tests cost more in terms of both the hire of testing spaces and the invigilator time required. Timetabling may also be facilitated with fixed test durations. There has also been an argument from the validity perspective, that standardised testing conditions for all test takers promotes fair assessment and valid interpretation of test scores[5]. For these reasons, time limits may be desirable, and most test developers and academics take a more balanced approach, setting time limits but trying to limit the speededness that results by making the time sufficient for most test takers. This has led to extensive efforts to analyse and measure test speededness in the US, as detailed in the book by Margolis and Feinberg (2020). Some of this research is detailed later in this report.

[5] Note that more recent theorising has moved towards the view that valid measurement should accommodate diversity (Sireci, 2020); in other words, a move towards more personalised assessment.

However, the speed/power distinction is not enough to fully categorise the full range of assessments, which may include some that are clearly not speed tests, but do set out to explicitly include the speed at which items or tasks are completed as part of the assessment. The following sections set out what should be a more helpful way of thinking about time limits in all types of educational assessments offered in England.

When speed may be a valid part of a test

In formal test theory from the US, speed tests are designed to measure speed of responding and are often psychological processing or perception tests where a fast, automatic process is tested. The aim is to measure how many items are responded to rather than the correctness of the response, assuming the individual items are easy. However, it is not just low-level psychological processes such as perception for which speed may be a primary measure.

The dual-process theory of cognition, which describes 2 distinct modes of thinking, is a long-standing theory in psychology. Many authors have offered different examples of this theory, and although there has been criticism of some of the details, they remain broadly accepted (see, for example, Evans and Stanovich, 2013, for a discussion of the debate). One description of this dual-process theory is given by Daniel Kahneman, in his book ‘Thinking Fast and Slow’ (Kahneman, 2011). We will use his terminology, since this book has helped to spread these ideas into mainstream culture. System 1 thinking is fast, automatic and unconscious, while System 2 thinking is slow, effortful and accessible to consciousness. Larger tasks may involve elements falling into either System 1 or System 2, and although the systems are distinct, they may interact. In some cases, a particular process may start as System 2, requiring effortful concentration and conscious thought processes, but as it becomes well practised and the individual becomes more expert at carrying it out, elements of the larger process may cease to require conscious control, becoming automatic System 1 elements. The more expert an individual becomes at a task, the more elements become automated, until eventually the whole process may fall under System 1, in other words carried out without conscious thought.

This migration is a part of the concept of fluency and expertise that is described below. A novice may carry out a task using System 2 thinking, but as expertise builds the effort required decreases until it may become relatively effortless System 1 thinking. For example, a novice typist needs to think where each letter occurs on a keyboard and measurement of their typing speed is assessing System 2 thinking. Once some level of typing proficiency has been reached, where conscious thought about where particular keys are located is not required, a test of typing speed would test System 1. Some quite complex cognitive processes can be automated, such as the ability of expert chess players to unconsciously and rapidly recognise and process particular arrangements of pieces and sequences of moves that allow them to more efficiently derive an appropriate response (de Groot, 1966, Chase and Simon, 1973).

The above are examples of fluency. Skills-based tasks may focus on fluency. Fluency has various definitions that apply to different contexts (for example in second language proficiency). But a widely cited work by Binder (1996) describes behavioural fluency as “that combination of accuracy plus speed of responding that enables competent individuals to function effectively in their natural environments” (page 163). This indicates the progression beyond an effortful and conscious approach to a task, where a level of proficiency is reached beyond mere accuracy in completing the task, but in which the task is completed accurately, more quickly and with less mental effort. This often indicates a degree of automaticity in the learnt skill, as described in the migration from System 2 to System 1. Fluency is a feature of expertise in many skill areas.

Knowledge fluency is also of relevance to educational assessments. The ability to efficiently access stored knowledge, or cognitive techniques, is key to problem-solving and is a feature of expertise in more knowledge-rich contexts (for example, Chi, Glaser and Rees, 1982). Those whose learning is deepest and most effective will have richly elaborated knowledge schemas – the structures in memory that organise and enable access to information. This is not just about providing rapid recall of knowledge, it is also about the quality of the information they store and the rich links (associations) to related knowledge. Such elaborate memory structures may enable more rapid solutions to complex cognitive tasks. For example, rich knowledge structures in maths may allow experts to recognise instantly that a particular problem requires solving with simultaneous equations. This would be an indication of expertise in maths.

In most fields, part of being an expert is the ability to carry out tasks and activities more quickly than someone less competent, so for a test designed to assess expertise, speed of response may well be construct-relevant. Lovett (2020) discusses timing issues around skills-based tests and different levels of expertise. Lovett notes that applying time pressure within a test emphasises the benefit of a high level of expertise. In some of the skill areas reviewed, accuracy and speed increase together as expertise builds, while in other cases, accuracy may plateau, but speed of working continues to improve as expertise increases.

In practice, in performance assessments that contribute to occupationally-related qualifications, speed may be a valid part of the test construct where it is an element of job roles the qualification is intended to train or qualify an individual for. The ability to perform practical tasks at speed in tests, and therefore to show a level of skills expertise, would evidence the ability to work efficiently and cost-effectively which would be valuable to employees and employers in that occupational area. It may be more usual for such tests to fall under the category described in the next section, where time limits are set to assure an adequate speed of working, perhaps commensurate with an appropriate level of competence. However, they may sometimes be assessed such that the raw speed at which tasks are completed is credited – faster is better, and leads to higher test outcomes.

There are also a variety of tests that set out to explicitly measure speed of performance within the more effortful and conscious System 2 thinking in Kahneman's (2011) framework. For example, aptitude tests used for some educational or employment selection purposes may look similar to academic tests since they may assess skills such as problem solving, language processing or verbal or numerical reasoning, but these may have quite strict time limits. The intention is to not just measure the ability of individuals to complete (sometimes complex) mental tasks, but to see how many they can complete in a set time. They include speed as a design element so that they can test the suitability of individuals for jobs where they would be expected to work accurately at speed, or at least where this skill is valued and considered to be relevant to candidate selection. Of course, not all selection tests are intended to be speeded, and the precise purpose of the test should be used to decide whether time should be factor in responding and scoring.

When completing tasks within time limits may be a valid part of a test

Some tests may set a time limit for completion which intentionally puts test takers under time pressure, but the test is not about raw speed, just working at an adequate speed. There is no additional credit for completing the task(s) much faster than required by the time limit, or indeed completing more tasks. These tests tend to be more practical in nature, rather than academic. They are about a demonstration of adequate speed on the task, often related to the speed at which a competent, but perhaps not expert, individual would perform. So rather than measuring the amount of fluency or expertise achieved, they measure whether a particular threshold has been reached.

Such an approach is common in assessments for occupation-related vocational and technical qualifications or apprenticeships. Both skills and knowledge fluency may be tested. Many practical activities such as constructing products or artefacts may be timed. Although absolute speed is not necessarily assessed, being able to work at a reasonable speed and thus demonstrate adequate expertise is important. An incomplete task will lose marks or potentially lead to a fail on the assessment.

Other tasks in assessments may mix practical skills with knowledge-based problem solving, such as in automotive repair. These are not just tests of dexterity or manual skills, and may require quite advanced problem-solving. For example, diagnosing a fault in a combustion engine vehicle may require a sequence of non-trivial tests to be carried out to narrow down the source of a fault, which could occur for mechanical, electronic, fuel supply or engine cooling reasons. The outcomes of each test need to be interpreted correctly in order to decide the next problem-solving step, and quick access to well-consolidated knowledge will help with carrying out the task efficiently. Other tests related to technical occupations may include tests that are almost entirely knowledge-based and require fewer practical skills. But again, completing complex multi-stage problem-solving activities under tight time limits can demonstrate a sufficient degree of expertise has been achieved to progress into that occupation.

However, some types of time-limit assessments may be more academic in nature, but still require clear time limits to make them authentic and therefore valid. This is particularly true of some license-to-practice tests or other entry tests for careers where rapid knowledge-based decision-making can be a factor. Objective Structured Clinical Examinations (OSCEs) for entry into medical professions operate under strict time limits. Candidates (such as student doctors) move from one ‘station’ to the next according to a fixed schedule. At each station, they are required to complete a medical procedure, diagnosis or consultation within a set time and are observed and assessed as they do so. As a reflection of the need for medical professionals to make diagnostic and treatment decisions rapidly, rather than having time to ruminate, time is again a core part of the construct in this test[6]. Note that time is of the essence in most medical cases, not just emergencies. In the UK, general practitioners are usually allowed only 10 minutes for each patient appointment, including diagnosis and prescribing treatment.

Margolis, von Davier and Clauser (2020) describe research looking at the equivalent practical exams in the US which are a part of the United States Medical Licensing Examination (USMLE). One study (Clauser, Margolis, & Clauser, 2017) manipulated the timing of one of the computer-based case simulations used in this assessment. Because one of the 9 simulation slots was not scored, they were able to set 3 different time limits of 15, 20 and 25 minutes (the standard time) for this slot, and they also varied the task itself that occupied the slot. They found that reducing the time available to 20 minutes did not have a large effect on scores for most tasks, but a 15-minute time limit did. This suggests that the 25-minute time limit might be appropriate if the intention was to apply nothing more than a moderate amount of time pressure. A shorter time limit might have been appropriate to generate more serious time pressure.

[6] Although extra time can be granted in OSCEs, the General Medical Council states that this can include extra question reading time, but extended station time only “at station assessments that do not directly replicate clinical practice.”

Speed of working, and thus fluency, is considered to be a valid part of the construct in other types of medical licensing. In the context of a licensing exam for osteopaths (the COMLEX), Lovett (2020) describes unpublished research in which a survey of faculty members at osteopathic medical schools showed that 93% agreed that knowledge fluency, defined as "the ability to recall and apply information accurately and without hesitation”, was a critical competency (NBOME, 2013). Eighty-five per cent also agreed that test time limits for the licensing exam should not be removed. Lovett also describes similar views among US lawyers regarding the need to perform under time pressure in bar exams for entry into the law profession. Respondents strongly supported for the need to read, write and think under time constraints, suggesting that speed was considered to be a construct-relevant part of the bar exam test.

As Kane (2020) describes, performance tasks such as those described above have a set time limit, so completing these tasks within this limit is important, but measured speed per se does not usually contribute to scores. In other words, completion time is not recorded and scored. Credit is generally given for the effectiveness of the performance providing it is completed within the set time limit, not the raw speed of the performance. In many cases this reflects the real-life context the test is designed to assess, where adequate rate of performance is valued, but absolute speed is not. So these tests require timeliness in the response rather than pure speed. He refers to these as time-sensitive performance tests, distinguishing them from tests of speed. The decision as to whether or not to record how long a test taker needed to complete a test and to factor this into their test scores is a context-dependent decision for test developers and test users to make.

When speed may not be a valid part of a test

Tests of knowledge, skills and understanding such as exams in GCSEs, AS and A levels do not usually include speed of working as a part of the test construct. As noted earlier, nowhere is the speed at which individuals can complete tasks under exam conditions in GCSEs, AS and A levels explicitly stated as a design requirement in any published documents. Similarly, the theoretical conception of power tests in the US holds that such tests should not be time limited so that they only measure knowledge and understanding. According to this view, knowledge-based tests should either be untimed, or the time limit set should be sufficiently generous not to restrict the ability of test takers to show their best performance.

Tests of practical skills are also not always intended to measure speed in the way that the fluency examples given in previous sections are. It may be solely the ability to complete a task (no matter the time taken), perhaps including an evaluation of quality, that is the measurement aim of the test. In such cases, a time limit may be set that is designed to be generous, or no time limit may be used. In practice, it is not common for assessments in England to be untimed because of scheduling requirements.

As discussed in detail in Kane (2020), if speed is not an intended part of the assessment construct, but tests are time-limited such that a significant number of test takers are not able to demonstrate their best performance in the available time, the ultimate ability of those time-pressured test takers will be underestimated. This would be a threat to the validity of the test outcomes. However, as Kane points out, there is probably, in most instances, an implicit acceptance that test takers should be able to complete the tasks set in “some reasonable length of time” (Kane, 2020, page 20). This suggests that in many cases there may be a view that an adequate speed of working is a worthwhile attribute for a test taker to have.

Extending this view a little, it is possible to ask the question as to whether in any knowledge-based test, knowledge fluency should be valued and measured. Is efficient access to knowledge, and deployment of this knowledge to complete a task quickly, a legitimate part of a test construct? The following sections consider this issue with a focus on exams in GCSEs, AS and A levels in England.

Should or could speed be a valid part of the construct in knowledge-based tests?

Knowledge fluency appears to be considered a valid element of some medical licensing tests, for example. It is, therefore, legitimate to ask whether it could be considered a valid aspect to include as part of the assessment in tests for subject-based general qualifications. For example, fluency in maths could be indicated by rapid selection of the right approach to a question, with direct progress towards the solution without deviation or missteps. The same view would suggest that the ease and fluency with which an individual accesses richly connected historical knowledge they have learnt when considering complex questions in history might also be related to knowledge fluency. Here, rapid selection of the right knowledge, and only the right knowledge, needed to answer a question could indicate well-structured and consolidated knowledge in memory, and expertise in historical analysis techniques.

Therefore, any academic subject-based test may contain elements of fluency. Indeed, in selected-response questions where there is no need to write anything, it might not be unreasonable to argue that speed of working could be a valid part of the test construct indicating expertise in the subject. However, there are 2 considerations, the first of which is of particular relevance in England, that might make speed a difficult element to integrate into the construct of such tests. The first is that for any knowledge-based test that requires constructed-response answers where a significant part of the time spent on a question is spent physically producing the response, writing or typing speed may be a major construct-irrelevant issue. The second is that in contrast to the discussion of fluency above, and despite the appeal of the idea that the most able (or best prepared) individuals will score most highly but also work through questions quickest, speed and ability tend not to be that closely related on exam-like tests when carefully analysed. Evidence looking at this relationship is described in detail in a later section.

Constructed-response questions on tests

Constructed-response questions used in exams for GCSEs, AS and A levels are varied in the type of responses they require. Some questions may require significant thinking time but only a short answer. Other questions may require both planning and writing time. Sometimes most of the total response time will be spent writing, although this cannot always be disentangled from planning time, since planning later parts of the response may be happening alongside producing an earlier part (in other words planning as you go). In some cases, an individual's handwriting (or typing) speed may be a limiting factor in the length of response that can be given in the available time. This may constrain their ability to maximise their performance.

It is probably more feasible to include fluency as an element of the assessment construct for tests that consist of short-response items, where outputting the response once it has been reached does not take long. Questions that require a word or short phrase in response may also lend themselves to measuring knowledge fluency as multiple-choice or other selected-response questions. Similarly, a mathematical problem-type question on a science paper, where the speed at which the problem space is formed in the mind and the speed at which calculations occur, might also lend itself to measures of fluency. Such questions are more similar to multiple-choice questions, in terms of the proportion of the total response time spent producing the response being low.

The next question is whether the ability to provide a lot of evidence of knowledge and understanding in a limited time should indeed be valued for exams in GCSEs, AS and A levels even if it could be included in the measurement construct. This is about the purpose of these qualifications. If they are considered to be about mastery of the qualification content, an argument would need to be made that this mastery includes an element of fluency. Thought would also need to be given to how this could be implemented in practice.

Another consideration may be whether the purpose of these qualifications includes predicting the future performance of an individual in carrying out related tasks. Would such knowledge fluency be important in the kind of work or other educational opportunities the qualification may allow access to? This is a question for all those involved in specifying and designing such tests, including any end users who may influence or directly contribute to design decisions. It may be unusual for most professions involving science or maths, for example, to demand this of practitioners in their work. Usually sustained and extended problem-solving and analysis involving careful thought and high levels of care and accuracy are likely to be valued, rather than rapid-fire knowledge recall. Knowledge fluency may be a small element of the whole process for a work task. That is not to say that a case could not be made that such knowledge fluency has value in some further opportunities those taking general qualifications might pursue. Time pressure in work situations can be highly variable and there will be some circumstances where speed of working (or thinking) may be critical.

Turning to tests consisting of extended-response questions, particularly full essays, technically it becomes harder to say that speed of working could be integrated into the measurement. This is because the largest part of the total question response time may be the writing or typing time. Unless writing or typing speed is part of the construct to be assessed, this poses a problem, and ideally the rate at which someone can physically produce a response of the appropriate length should not be a part of measurement. Such tests should therefore not be speeded.

However, even here, there may be some arguments concerning the positive value of time limits in essay-based tests. Some jobs involving writing do have tight deadlines, for example journalism, and many involve the need to write summaries or briefings quickly and concisely to inform others. Therefore, an argument could be made regarding responding to extended response questions, where answers that address the given task consistently, with concision and a strong structure could be considered better, or be more highly valued, than longer responses that cover identical rewardable evidence but mix in irrelevant or even wrong evidence in a less focussed response. Although this is not speed of working per se, it could reasonably be argued that relevance, clear structure and concision in response are valuable skills for life. Indeed, the example of journalism given above is one where concision in extended writing is very highly valued.

Time limits will often serve to stop test takers writing as much as possible – the ‘get everything you know down in case it’s relevant’ type of response. It is probably not uncommon for individuals to feel that the more they write in response to essay questions the better. This may be because by doing so they believe they increase their chance of including rewardable content, especially given that mark schemes in England typically reward students positively for relevant material, rather than penalising irrelevant material. This may be particularly true for someone who is not sure what is relevant material to include in their answer. Some test takers may simply believe more is better, and that they may impress the marker with their breadth of knowledge.

However, research shows that in practice, longer is not always better. Benton (2017) looked at marks awarded to GCSE English literature essays of different lengths. For a 45-minute exam paper requiring a single essay response, average scores increased with essay length up to around 700 words, but beyond this word count, average marks plateaued, and sometimes showed a hint of dropping. Overly short responses do limit scope to demonstrate what the question is assessing, so that up to a point, longer responses provide scope to obtain more marks. But once the response is long enough to have sufficient material to demonstrate all of the knowledge, skills and understanding captured in the mark scheme, adding more words may add no marks, or in some instances could actually reduce marks by making the work appear less well organized or by including a larger number of negative features (for example mistakes or misspellings). Indeed, fatigue when writing very long responses may lead to a decline in quality as the response progresses.

Unfocussed and overly long responses can also cause problems in marking. They increase the time required to mark responses, and can also reduce the reliability of marking, since examiners may find inconsistent responses (good mixed with bad elements) harder to evaluate, and different examiners may factor wrong or irrelevant information into their marking judgements differently (these are one type of ‘hard to mark’ responses – see Morin, Black, Howard and Holmes, 2018). Therefore, time limits in essay-based subjects may be somewhat helpful in focussing answers, and administering them as pure untimed (or very generously timed) power tests could lead to over-long, unfocussed answers.

At this point it is important to note that time limits are just one way to encourage concision or well-structured answers of the appropriate length. Word limits or response area limits could be used instead of time limits to discourage over-long responses. Given the level of test-wiseness in students today, both the size of response areas and the marks available for questions typically provide a fair indication of the length and depth of response expected.

While there may be arguments that some aspects of the process for answering constructed-response questions may validly include speed as a desirable feature, more work would be needed to determine precisely which features these might be. Analysing the different stages required to answer a question, and categorising them on whether the time they take is indicative of proficiency or fluency, would be helpful. This could feed into further discussions about which types of questions and tasks may best lend themselves to having strict time limits, if speed of responding was to be valued. A final consideration may be whether such fluency or proficiency should necessarily be expected of a 16-year-old, or even an 18-year-old, taking GCSEs or A levels, or any other test.

If speeded, how speeded should tests be?

If it was decided that speed of working should indeed be part of the construct for some (or all) exams in GCSEs, AS and A levels, this would need to be made explicit. For tests on paper, this would need to be implemented at the level of a whole test, with tight time limits for the test forcing a reasonable speed of working, which would therefore penalise those who work slower because they will be rushed on some items or they will not have time to attempt all items. The alternative of collecting item timing data to feed into scoring, or limiting item-level response time, would not be feasible without on-screen assessment. From the perspective of fairness to test takers, the approach would need to be clearly articulated and well supported so that all test takers would understand how and why they would be assessed in this way. Indeed, strong justification would need to be made to support such an approach. If speed were to be part of the assessment (in terms of intentionally tight test time limits) this might mean that extra time would no longer be a valid examination adjustment to offer if completing the tasks in a fixed time was the design intention. This issue is considered further in the discussion section of this report.

The purpose of the test and the way speed of working is integrated into the test construct would need to be stated clearly. There would, therefore, be significant choices to be made regarding just how speeded tests should be, and whether this should vary across subjects and qualifications, considering some of the points raised about question types in the previous section. For example, assuming speededness could be measured fairly precisely, would 50% of students experiencing a test as speeded be about right? Or should it be far fewer, perhaps only a few percent, or almost none at all? The effect of test speededness will be very different for each test taker (and indeed, potentially for each cohort taking a test). Some may feel no impact from the test time limits and be able to tackle each question to the best of their ability. Other test takers might have been able to achieve significantly higher scores with more time. So in deciding how tight time limits should be, how much should scores be impacted in the most severe cases?

While metrics have been suggested to allow judgements to be made of whether a test is speeded (for example Swineford, 1956, described later), they are entirely arbitrary. It could be argued that if one individual is impacted by one mark then a test is speeded. Perhaps most people would agree that this would not be sufficient evidence to judge that a test should be labelled as speeded overall. But there is no clear consensus on the proportion of test takers affected or the size of the impact on scores for a test to be judged speeded. Any thresholds set are likely to be arbitrary and so it may prove to be very difficult to arrive at a consensus for what is appropriate speededness, even if it were to be argued that speed of working was a desirable element to measure.

The potential benefits of time limits

Setting aside the question of whether speed of working should be directly assessed, it is also worth noting that having time limits for tests can have some benefits beyond just logistical convenience. Firstly, when time limits promote the ‘right’ amount of time pressure, this may have a positive influence on performance through motivation and physiological arousal effects. Conscious motivation to perform may be maintained when a degree of time pressure is perceived to be present, particularly where the exam is not so long that boredom or maintenance of concentration is a challenge. A known, constrained, test duration may also help test takers to prepare themselves beforehand, and to maintain concentration throughout the test.

Physiological arousal is also an important component of human performance. Arousal refers to the nervous system’s state of activity and readiness for action. In the most extreme cases this is exemplified by the ‘fight or flight’ response – clearly a higher state of arousal than is preferable for taking tests. Although there have been some controversies over the application of the Yerkes and Dodson (1908) performance-arousal curve to seemingly every aspect of human performance, there does appear to be an inverted-U relationship in many types of tasks such that performance is optimal with sufficient, but not too much, arousal. When considering exams, this arousal will be caused by both the testing situation and its stakes, and the perceived time pressure.

It is not clear exactly what the ‘right’ amount of arousal would be in exams, and this would certainly differ across individuals. It may also be that for some or many individuals, simply being tested (regardless of time pressure) would generate sufficient arousal because of the stakes of the test (assuming that the test is high stakes). For some, arousal might be too high, leading to test anxiety (see, for example, Putwain and Daly, 2014), and perceived time pressure from time limits may contribute to this. Too much anxiety is presumed to reduce test performance for some individuals, likely countering any positive motivational effects (however, see Jerrim, 2023 for evidence that reported test anxiety may not necessarily affect exam outcomes).

Time limits can also be helpful, to a degree, in setting expectations for test takers when the test includes constructed-response items. As noted earlier, appropriate time limits may help to signify what an appropriate length of response should be, such that candidates should have time to produce valid evidence, but not time to write everything they know including irrelevant information. However, because less well-prepared or knowledgeable test takers may often struggle to respond with only relevant material, and so may want to write too much, they may still perceive that the available time is too low.

Research into the relationship between speed and ability

The understanding of fluency and expertise described earlier suggests that in general, the more proficient the individual, the quicker they would be at completing a test that assessed their ability. For skills-based tests, there is a lot of evidence that speed on many of these tasks increases as expertise grows. Knowledge-fluent individuals can access the right information rapidly due to the richness of their knowledge storage in long-term memory, and are also quick to identify and select appropriate and effective problem-solving approaches. We might also expect these individuals to be faster at completing knowledge-based tests. The most able or best prepared test taker should be able to answer questions quickly and accurately and so score highly, but also complete tests quickly. The less able or less well prepared should struggle more on questions and would be expected to take longer to access the knowledge required, or to work through the task presented. This means that they will score lower and take longer to complete a test, or indeed, might not finish it.

As noted earlier, this might be especially true of tests consisting of selected-response questions, where outputting the response is not a significant part of answering a question. Instead, accessing knowledge and thinking through any required mental steps to reach an answer may be the most time consuming element of the response. The relationship may be less direct for tests where, for example, reading time (when long written sources or extracts are used) or handwriting speed are more significant factors.

A belief that ‘smarter is faster’ underpinned early thinking about test completion and time-on-task in multiple-choice testing in the US. Early theorising (for example Spearman, 1927) suggested that speed of responding and response quality were closely related, and that those who responded more accurately would also be quickest. On this basis, because the main aim of the tests described was simply to rank order test takers rather than measure absolute levels of performance, strict time limits would not disrupt this rank order as the most able would still complete the most items even when they did not finish a test. Some early evidence was put forward suggesting that this assumption was correct.

However, Davidson and Carroll (1945) criticised the methodology used of comparing scores for each test taker at the end of a timed part of a test and their scores following a further additional untimed period. They pointed out that the untimed score included the timed score, and so the correlation between the 2 would be inflated. Similarly, inflated correlations would arise if the tests examined were actually not very speeded. In the tests described, it was unclear how speeded they actually were.

In many cases the timed score might have been identical to the untimed score if the test taker completed the test within standard time.

Davidson and Carroll carried out their own study, using factor analysis to determine the factors that explained how scores on several different tests varied at the end of standard test time, and also following a further untimed period. They also recorded the time at which each test taker finished the test. They found that performance following the additional untimed period was explained by an entirely separate factor (technically, in factor analysis terms, an orthogonal factor) to the time taken to complete the test, meaning that performance and speed were uncorrelated across test takers. They also found that scores at the end of the timed part of the test were explained by both speed and power factors, suggesting that the scores were a combination of the 2 factors. As quoted earlier, Davidson and Carroll were highly critical of treating scores on timed tests as representing only power (in other words ability).

Further work (for example Mollenkopf, 1950; Tate, 1948) supported Davidson and Carroll's conclusion that speed of working and ability were not closely correlated in knowledge-based tests. Researchers had also suggested that speed of working was a consistent trait of test takers, independent of their proficiency, and determined by a variety of individual characteristics (Himmelweit, 1946; Kennedy, 1930; Tate, 1948). Evidence has continued to accumulate that speed and ability on academic tests are independent, despite the superficial appeal of a link. For example, Gernsbacher, Soicher and Becker-Blease (2020), who recently argued against the use of time-limit tests, review a number of studies over the last century that show little or no predictive link between speed and attainment on a variety of academic tests across many subjects and educational levels.

The relationship, or seeming lack thereof, between speed and ability is likely to be influenced by the fact that in answering questions or completing tasks, fluency is only a part of the overall speed of working. The pace adopted by test takers may not be their maximum pace, especially when the test time limits do not put them under a large amount of time pressure. They may dedicate variable time to reading, understanding and conceptualising the question, planning an answer (if a written response is required), confirming answers they have just reached by working through questions again, and finally checking answers at the end once all items have been attempted. In cases where a test is highly speeded and individuals have to work through questions at or near their maximum pace, knowledge fluency may be a somewhat more significant factor, especially in testing with selected-response questions.

Decisions on the speed of working test takers adopt may be quite individual, and based as much on character traits such as confidence as on ability. The speed/accuracy trade-off describes the process thought to underlie these decisions.

The speed/accuracy trade-off

Individuals will generally perform better or be more accurate on individual tasks (for example exam questions) if they take more time on them, but will complete more items with lower accuracy if they take less time. This speed/accuracy trade-off has been well studied in psychology (for example, Heitz, 2014; Luce, 1986; Wickelgren, 1977). It can be traced all the way back to the early days of experimental psychology (for example, Henmon, 1911) when the trade-off was apparent in a variety of perceptual and cognitive tasks.

Early theorising by Thorndike, Bregman, Cobb et al (1926) and Thurstone (1937) also factored item (or task) difficulty into the speed/accuracy trade-off. Thurstone's conception suggested 3 consequences of the trade-off. First, for a fixed time, the probability of a correct answer decreases with increasing item difficulty. Second, the probability of a correct answer increases with a longer response time. And third, with increasing item difficulty, more time is required to respond correctly. In a time-limited test, each test taker therefore needs to make an individual judgement on the time to spend on items based on their own judged probability of success on that item.

The speed/accuracy trade-off can be applied to academic tests of knowledge, skills and understanding where time limits mean that test takers need to work faster than they might if time was unlimited. Test takers may need to decide their pacing to determine how much to sacrifice accuracy to try and finish the test, or to accept that they will not finish and concentrate on accuracy and depth of response on those questions they do answer. Perseverance on items that an individual finds difficult, but potentially achievable, also becomes a factor. The rest of this section considers some research into the factors that influence decisions about time allocation to items. This US research is largely in the context of tests composed of one-mark multiple-choice questions.

Allocation of time to items and accuracy in time-limited tests

Research shows that test takers can be strategic in the way they allocate time to items depending on the judged time pressure in the test as well as their own ability to answer questions of varying difficulty. Harik, Feinberg and Clauser (2020) reviewed a body of research considering item difficulty and response time in the context of a high-stakes multiple-choice knowledge test used for medical licensing (the USMLE, referred to earlier). The test is completed online using a system that collects full item response-time data. Although the sections of the test analysed were non-scored experimental sections, these were embedded in, and largely indistinguishable from, the scored operational sections so test takers should have been highly motivated to perform.

Harik, Clauser, Grobovsky et al (2018) showed evidence of careful allocation of time to items when different length versions of the same test were presented with the same time limits. Selective removal of items created different versions of the hour-long test section with 28 to the full 44 items. Right from the start of the test, item response time varied according to the number of items in the test, with less time allocated per item for the longer test forms. Test takers appeared to demonstrate some understanding of how much time to allocate to each item in the different-length versions of the tests to give them an opportunity to complete all of the items on the test.

This time per item was generally maintained throughout the test, with only a slight decrease in time per item towards the end for the shorter (and therefore more generously timed) test forms. For the longer test forms, a larger decrease in time per item was seen towards the end of the test, indicating that the pace adopted at the start was insufficient to complete the test and so an increase in pace became necessary towards the end. This suggests that test takers attempt to set an appropriate pace (selecting a point along the speed/accuracy continuum) right from the start of the test, but in highly time-limited conditions they either underestimate the required pace or cannot reach it.

The speed/accuracy trade-off suggests that performance should be lower when a higher pace is adopted. Several studies support this prediction. A large-scale study by Bridgeman, Trapani and Curley (2004) looked at performance on a pilot section of the SAT test not used for grading. Like Harik et al (2018) they created versions of this test section with different numbers of questions that were taken under the same time limits, by deleting carefully selected questions to maintain equivalent difficulty across the test versions. The results showed that, for items throughout the test, average scores were lower under the less generous timing condition, not just the items at the end. The finding that time limits affected performance throughout a test were broadly replicated for computer administration of the ACT by Li, Yi and Harris (2016). The analysis by Harik et al (2018) also showed this, with performance throughout the test a little lower for the most time-limited test forms. The most substantial drops in performance were seen towards the end of the longer test forms, where time per item was most reduced.

Harik et al (2018) also found differences in the time allocated per item for the 2 shortest and least time-limited test forms, but which was not reflected in any difference in accuracy of response. This suggests that when more time was available in the shortest form, test takers took more time than was needed to maximise their score. Related findings were reported by Wise (2015), who looked at time per item through a large-scale K-12 (US elementary through to secondary school) achievement testing program. He observed that for the first 10 or so items on a 50-item test, test takers took much more time than they did through the rest of the test, but accuracy remained the same throughout the test. They appeared to be taking more time than needed to maximise performance early on in the test.

It would appear then that when the time allowance for a test is generous, test takers may fill the available time, pacing themselves more slowly, or they may complete the test and spend more time checking and reviewing answers. Indeed, Harik, Feinberg and Clauser (2020) analysed patterns of item review in USMLE data and found an increasing amount of time was spent returning to items in the shorter test forms, even where this checking did not increase scores. This indicated that test takers had already answered all questions as well as they could but continued to double-check their answers. Therefore, it appears that any measure of speededness based on amount of time test takers need to complete a test, or the number of test takers making use of all available test time, may overestimate test speededness.

Harik et al (2018) also looked for evidence of rapid-guessing behaviour. A small amount of rapid guessing (in the form of very rapid response times) was seen throughout the test, and this increased towards the end. Unsurprisingly, more rapid guessing behaviour, as well as item omission, was found in the longer test forms. However, although this partly explained the drop in performance on items towards the end of the test, it could not fully explain it. This indicated there were some responses where the item had been attempted, but hurried – a different point on the speed/accuracy continuum had been selected. The lower average time per item towards the end of the test (and corresponding lower accuracy) was therefore a mixture of carefully considered responses, hurried responses and rapid guessing.

Test taker proficiency and item difficulty

Speededness of the type of multiple-choice test frequently investigated in the US literature appears to be partly related to the proficiency of the test taker. In further analysis of the USMLE test data described in the previous section, Harik et al (2018) stratified the test takers by proficiency (based on their overall score). They found only moderate differences in overall score across different-length test forms for the highest proficiency group, suggesting that they did not find even the longer versions very speeded, but larger differences across test forms for the lower proficiency groups, suggesting that these individuals found the test quite speeded.

Considering the allocation of time to items, a study by Swanson, Case, Ripkey, Clauser, and Holtman (2001) modelled data from the USMLE. They found that average response time was higher for more difficult items as Thurstone predicted, but there was an interaction with test-taker proficiency. Lower-scoring individuals spent more time than higher scoring individuals on easier items, and less time than higher-scoring individuals on more difficult items. Harik, Feinberg and Clauser (2020) repeated the analysis of Harik et al (2018) on a similar dataset from USMLE trialling with different length experimental test sections. They found the same pattern as Swanson et al, with lower proficiency individuals allocating more time to easier items than higher proficiency individuals and vice versa. This pattern was more evident in the more generously timed test forms, probably because this allowed scope for individuals to persevere on items they found difficult.

Because of this proficiency dependence, complex relationships between item response time and accuracy may emerge when individual items are considered. Feinberg and Jurich (2018) analysed the entire cohort taking the USMLE test. While the speed/accuracy trade-off suggests that for a single person, spending more time on an item increases response accuracy, for each individual item they found a non-linear relationship between the probability of a correct response and item response time across this mixed-ability cohort. The probability of giving a correct answer was low for very short response times (largely indicating rapid guessing), but this probability rose rapidly to a peak at a duration well below the average response time. Following this peak there was a gradual decline in probability correct as time spent on the item increased. At the same time, when Feinberg and Jurich looked at the distribution of aggregated response times for all the items on the test within different proficiency groups, there were only small differences and the most proficient were only slightly quicker in responding overall.

This latter finding may appear puzzling given how accuracy and item response time varies on single items. But these findings are consistent with the earlier finding that different test takers will spend the most time on different items based on their proficiency. Items that are either very easy or very difficult for that individual will have little time spent on them. But items that are found to be quite difficult, and may or may not be possible for the individual to answer, will have the most time spent on them. Therefore, for each individual, more time will be spent as items get more difficult, up to a peak, after which increasingly less time will be spent as the items become increasingly less answerable. The difficulty level at which peak response time occurs will vary by individual ability.

These effects, especially the reduced time spent on difficult items, will be partially based on test takers' ability to accurately assess their likelihood of success. It is also important to remember that the test must be speeded to an extent for these findings to emerge – otherwise a motivated test taker might simply spend more time on the increasingly difficult items as Thurstone (1937) suggested. In a time-limited situation, and where the test taker is prepared and does not behave in a naïve, simplistic way, careful allocation of time based upon probability of success per item should emerge. Certainly, the USMLE would be a test that individuals would be well-prepared for since it provides the entry point to a medical career. Wider issues of test wiseness and preparation are considered further in a later section.

Further support for the ability-dependence of time allocation comes from statistical models fitted to wider test data. Hierarchical item-response theory (IRT) models (described more fully) such as those described by van der Linden (2007) model the relationship between response accuracy and response time. In these models, each test taker has an ability parameter based on item response accuracy and a separate speed parameter based on their response time per item. The relationship between speed and ability in the model can be positive, negative or flat, depending on the test, so sometimes more able test takers are faster, sometimes slower. Test takers are therefore modelled by 2 independent traits of speed and ability, capturing the distinction that had been made by researchers many years before (Himmelweit, 1946; Kennedy, 1930; Tate, 1948).

The detailed findings above regarding differential effects of ability and item difficulty are also captured in these models, with a slightly different pattern because they generally model a linear relationship between response time and accuracy. For example, while for most items there may be a relationship such that longer item response times predict lower accuracy (as in the tail of the pattern described by Feinberg and Jurich, 2018), for more difficult items the relationship may be non-existent or reversed, with longer response times related to more accurate responses. For each individual test taker, the general relationship that longer response times predict lower response accuracy also varies by ability. The relationship is typically stronger for high-ability test takers (so faster is more accurate) but this is absent or even reversed for low-ability test takers, meaning that slower may be no less accurate than faster for the less able (Bolsinova, Tijmstra and Molenaar, 2017; Goldhammer et al., 2014).

These relationships in the models described above are captured as linear functions. But all of these findings, and further deeper analysis of the modelled data, suggest a nonlinear relationship between speed and accuracy. Therefore, statistical modelling results (summarised in De Boeck and Rijmen, 2020) across a variety of test contexts are largely consistent with the experimental results described above from the USMLE. Longer response times are associated with lower response accuracy for more difficult items, and for higher ability individuals. For easier items and lower ability individuals, there may be no relationship, or longer response times may be associated with higher accuracy. All are consistent with the idea that test takers spend time on items depending on their ability and the item difficulty, and that sometimes this is a strategic decision based on probability of success.

When these findings are extended to consider how overall test speededness varies by test taker proficiency, the pattern will depend on the relationship between the difficulty of the set of items in the test and the proficiency of the test takers. If individuals with a range of proficiencies take a test with items that are well matched to their proficiency, it may be that both the least and most proficient find the test least speeded. The former can answer few items, while the latter can quickly and easily answer most items, with just a few items slightly beyond their ability that they persevere on. Moderate proficiency individuals may be under the most time pressure since there may be a significant number of items that they find difficult but achievable.

In support of this idea, in the Bridgeman, Trapani and Curley (2004) study described earlier, test takers were stratified into proficiency groups based on their scores on the operational parts of the test, and the largest drop in performance on the longer test forms was found for those within the moderate proficiency group. For the low or high proficiency groups similar performance was seen across the different length test forms.

This may not always be the case though. A mismatch in test taker proficiency and test difficulty could produce very different speededness outcomes. If the test is too hard for most test takers it will appear to be relatively unspeeded, since most test takers may experience low time pressure because they can only answer or attempt some questions – they are effectively knowledge-limited rather than time-limited. If a test is generally low difficulty for most test takers, it may only be the least proficient that experience the test as speeded. This appeared to be close to the pattern observed in the Harik et al (2018) analysis above. Therefore, because individuals will experience speededness differently, the proficiency of the cohort taking a test can make a big difference to measures of test speededness.

It should be borne in mind that these observations all apply mainly to tests composed of many multiple-choice items. Outputting responses (in other words, writing) is not a factor, and because the items all have the same value (one mark), choices around sequencing and time allocation across items with very different tariffs are absent. However, these findings on the way response time and response accuracy interact are important in showing that there is not generally a simple relationship between speed and ability. It is not necessarily the case that the most able are always faster in completing tests; rather, there is a complex relationship between the time spent on items, the item difficulty and the individual test taker’s proficiency.

Summary thoughts on assessing speed of working in written tests of knowledge, skills and understanding

Speed of responding is rarely (if ever) an explicit part of the assessment construct for more academic tests of knowledge, skills and understanding (this has been noted by, among others, Lu and Sireci, 2007). Since these tests are only intended to measure knowledge and understanding, test speededness should ideally be minimized. This would align with the ideas of “universal test design” (see AERA et al, 2014, Standard 3.1; Thompson, Johnstone and Thurlow, 2002) where tests are designed so that they are equally accessible for all, regardless of individual test-taker characteristics. The aim is to remove any barriers that are irrelevant to the construct being assessed, and so the need for test adjustments is minimised. Therefore, where speed of response is not part of the test construct, it is considered to be good practice to analyse test functioning to determine whether a test is indeed speeded, since sufficient time for all (or at least most) test takers should be allowed.

However, the idea of fluency and high proficiency in a knowledge area could imply that aspects of responding to a question or task might include the speed at which they are completed as a valid measure. Rapid recall of relevant information or efficient identification and selection of appropriate (or optimal) problem-solving techniques may all be considered signs of superior expertise. But how to integrate time constraints on certain types of tasks or even elements of certain tasks is unclear when for other elements of a response (most obviously handwriting) speed is clearly construct-irrelevant, and will make a noticeable contribution to response time for many item types.

Integrating speed of response as an explicit design intention would appear to be entirely impractical in current paper-based testing in England. Technology-based assessment, if it were adopted for GCSE, AS and A level exams in England, may offer some scope to include speed in elements of the response in test scores. This could be through restriction of the time available to give an answer (or complete parts of the response process), or by including response time in the scoring algorithm. However, even with technology, the fact that speed and ability do not closely align in many knowledge-based tests confuses the issue. Even in selected-response tests with numbers of 1-mark items, the relationship that the most able are also the fastest does not hold, because there are so many individual factors that may determine the speed at which an individual answers a question other than proficiency on the test construct (Tate, 1948).

Each response requires a threshold judgement to be made by the test taker: that the response is sufficiently complete (if written) or sufficiently likely to be correct. An individual’s propensity to check, double-check, return to items and amend or add to responses will all be influenced by factors other than ability, such as caution, confidence or other personality factors. Similarly, allocation of time to planning how to answer a question, particularly an extended-response question, is part of the speed of working. Some test takers may quickly start writing, while others may want to spend significant time carefully planning or structuring their response before beginning to write it. All of these individual factors together with the test features will feed into the overall speed of working that a test taker is most comfortable with or is capable of when under time pressure – in other words where they locate themselves on the speed/accuracy continuum.

Therefore, while it may be possible to restrict the time in which a whole response or part of a response is given, or to credit faster item-level responses in an assessment, this would entirely change the approach required. It may particularly disadvantage individuals who are not comfortable working at their very fastest possible speed even when they may be highly proficient. As noted earlier, exams can cause anxiety, and care would be needed if test takers were intentionally pushed far from their comfort zone in terms of how they had to pace themselves.

The research that is available on use of time in tests has been largely focussed on online multiple-choice testing, so is limited in its applicability to paper-based tests in GCSEs, AS and A levels. Although this research does show signs of careful thought by individuals in the way they used their time, such considerations will be more complex in test formats in England where constructed-response items with variable marks available are commonly used. The next section turns to time considerations that apply specifically to the testing context in England.

Time considerations in assessments in England

Before considering GCSEs, AS and A levels at length, we begin this section with a brief consideration of vocational skills-based testing. Fundamentally the assessment of work-based skills and performance does not differ significantly in England from any other jurisdiction, being essentially based on authentic work tasks. Whether speed of working on completing tasks is a desirable part of the test construct is a decision to be made by the assessment organisations or the professional bodies that these qualifications relate to. If speed of working should be part of the test, then time limits need to be set based on professional expertise, judging how long tasks should take based upon the proficiency level (and fluency) expected of test takers. Care should be taken that lime limits realistically reflect what is reasonable to expect of a borderline-proficient test taker, and not the norm for experienced professionals. Otherwise setting the right time limit for a performance test is a sector-specific judgement. In cases where speed of working is not required to be measured, almost certainly time limits will still be needed for practical reasons. But these should be set so that few, if any, test takers will run out of time if performing at a proficient level. Therefore, work-related skills testing does not raise specific difficulties, providing the duration of tasks is set based on reasonable professional judgement of the time required.

However, for GCSEs, AS and A levels in England that test subject-based knowledge, skills and understanding, the testing context is very different to that underpinning the theory and research of Margolis and Feinberg (2020). In England, testing takes place predominantly through written exams, which for most test takers are a physical paper and pen test[7]. Although some selected-response test items (generally multiple-choice questions) are used, most questions require a written response, ranging from a word (or number) to a sentence or two and up to extended responses, which might be multiple pages in length.

The theoretical knowledge and skills tests used in some vocational and technical qualifications, such as functional skills, and performance-table qualifications such as applied generals (including BTEC Nationals and Cambridge Technicals) take a very similar approach to GCSEs, AS and A levels in terms of item and response types. One difference is that some awarding organisations have implemented computer-based testing (CBT) for these qualifications, often in parallel with paper-based testing, although the test items are usually similar. The way that time limits are set for the tests and the expectation of the speed at which test takers work tend to be similar to tests in GCSEs, AS and A levels. Therefore, the following discussion regarding tests in GCSEs, AS and A levels also applies to these written exams for vocational and technical qualifications.

Many qualifications also contain assessed coursework, normally called internal assessment in vocational and technical qualifications or non-exam assessment in GCSEs, AS and A levels, that sit alongside exam-based testing. Ideally these qualification components should have no time limits, except where they need to be carried out under controlled conditions – in other words when they are directly supervised. Here, logistical considerations are likely to combine with judgements of reasonable task completion time to set the supervised-time allowance. Much like skills-based testing in which speed of working is not assessed, care needs to be taken that sufficient time is allowed so that individuals are not put under time pressure. In cases where students complete work outside of class, time limits can be used as guidance to specify reasonable expectations of the task to the students. Ultimately though, the time allowed to complete a task may only be defined by submission deadlines. Of course, students may still leave coursework until just before the deadline and thus put themselves under time pressure.

Returning to the examined components of GCSEs, AS and A levels, although these are intended to measure only knowledge, skills and understanding, in practice there

[7] As noted earlier, some students complete the tests by typing answers on a computer where this can be shown to be their normal way of working or as a reasonable adjustment or access arrangement to meet a disability or other access need. Other reasonable adjustments or access arrangements may include technological solutions or a human scribe to complete the test. However, the test itself is unchanged in these cases, only the mode of responding changes.

are good reasons why exams are carried out with fixed time limits. Although the costs associated with the use of paid-for testing centres in the US is not generally an issue in England since GCSEs, AS and A levels are mostly sat within school and college facilities, the scheduling of GCSE, AS and A level exams within the available time is an issue. As the period over which exams take place in the summer term is limited, there are multiple exam sittings (2 or 3) each day. Schools and colleges cannot easily run more parallel exams due to resource constraints such as space in which to seat test takers, or numbers of available invigilators. The pressure on resources also appears to have grown as more students receive access arrangements of all types for their exams.

The idea of untimed exams would be a practical impossibility in the current system of high-stakes summer exams in England. They would create scheduling issues given the need to fit in multiple exams per day in the summer exam series. Because the exam series takes place within a fixed window – to allow for teaching time in schools and colleges and to facilitate admissions to further and higher education – extending that window would involve significant challenges. Extending current exam durations would similarly present significant logistical difficulties. Longer exams would also be inconsistent with the recommendations on reducing the volume of assessment in the recent Curriculum and Assessment Review (DfE, 2025). Although the causal effect of exam stress on student wellbeing is not completely clear, exams such as GCSEs and A levels should not be tests of endurance for candidates. The ability to maintain concentration may be a valuable skill or attribute for life, but is not part of the assessment construct, so lengthening exams too much might not be ideal since it could introduce construct-irrelevant variance into test scores.

Setting time limits for tests

Time limits for exams in England appear to be set largely through experience and convention. Assessment professionals such as senior examiners working in awarding organisations who are involved in designing assessments play the primary role. Experience, and precedent such as the time limits set for previous exams, and knowledge of the way these functioned, is important. Questions may be asked such as, ‘does the general quality of response fall short of what test takers may reasonably be expected to demonstrate given the designed difficulty of the test items, and could this be related to the test timing?’ Attention might be paid to whether there are significant numbers of items not answered where expected difficulty would not predict this. Or is there evidence of a lot of incomplete answers, such as abruptly ended responses in longer-response questions that look more like time running out than knowledge or motivation running out? Such analyses are unlikely to be definitive.

Part of the question paper development process does include one or more scrutineers – subject experts – sitting the paper and answering the questions as a student would. Part of this process is a check that the paper can be completed in the time available (for example, AQA, n.d.), but given these are expert adults it is not clear that this is a particularly sensitive way of detecting overly tight time limits. Data on item completion is also available from live test administrations, and in ensuring that assessments are fit for purpose, awarding organisations should consider whether such data suggests test takers are running out of time.

Awarding organisations may also consider the views of stakeholders, particularly teachers and students. Frequent reports from schools and colleges that students are running out of time may influence decisions on time limits or test content when reviewing and revising assessments. Ofqual also has a role to play in looking at examination content and time limits – certainly for GCSEs, AS and A levels, where sample assessments are produced by exam boards as part of the qualification accreditation process. Part of the review of these materials is whether the time limits set are judged to be reasonable for the assessment tasks, again based on experience and comparable previous assessments.

The longer a test is, within reason, the more reliable are the outcomes. In other words, measurement precision goes up with more assessment evidence. This means that there is a tension between having sufficient evidence on which to evaluate test takers (increasing measurement precision) and limiting the length of assessment for a qualification to limit the burden on schools, colleges and test takers themselves. Alongside this, the qualification content, or syllabus, plays a part since sufficient content coverage in assessments is part of the validity judgement made by test developers and the regulator. A lot of content to assess may increase the pressure on the overall length of assessments, since there is an expectation that while each set of assessments for a qualification may only sample the content, each individual should be tested on sufficient breadth of content for the assessment to be valid. In addition, there is an expectation that in such a sampling model all content will be sampled at some point over a reasonable number of assessments.

At this point it is important to emphasise that even when the amount of assessment has been agreed in terms of numbers of questions and tariffs, this does not precisely define the time that should be allowed. Although there exist rules of thumb, such as ‘a mark a minute[8]’, even were this or some other rule to be followed rigidly, this would not necessarily define appropriate time limits. The time required to evidence one mark will vary by question type and subject, and the number of marks allocated

[8] This heuristic appears to be more relevant as guidance to help prepare test takers, rather than something embedded in the design of tests, although it would have no utility to test takers if it did not reflect, at least approximately, some reality about typical test design in England.

to a set task is an assessment design decision, not necessarily an intrinsic quality of the task itself. For example, the same essay question might reasonably be scored out of 8, or out of 50. The marks assigned to items are often just to signal the type and length of response expected, or to achieve the correct weighting for the item within a larger assessment.

Therefore, as noted, time limits are set more by precedent and experience than through any formal rules, with the intention that there is sufficient time for most test takers to complete the test. It is important to remember that without word limits imposed on responses (which would be hard to enforce in handwritten exam responses), and with variable writing speed across individuals anyway, it is hard to say in advance what a 'correct' time limit would be for a new written exam. There will always be a degree of imprecision in these judgements when new assessments are being designed and developed, and ultimately, given the variability across individuals, there may be no such thing as a correct time limit.

Piloting a new test before it goes live might provide some useful information, although it is unclear the last time any new exam in England was piloted in full with a view to investigating speededness. However, even such a pilot cannot be definitive since the conditions are likely to be perceived as low stakes by the test takers, limiting their effort. In addition, because of the way that test takers adjust the time they allocate to each item based on the available time limits, a pilot test administered under the standard time limits may underestimate the time required to maximise performance. The reverse may be true in an untimed administration, where test takers may take longer on each item than they need to achieve their maximum performance, leading to an overestimate of time required. How this interacts with the low test-taker motivation that may result from the low-stakes nature of a pilot test would be difficult to determine.

Extra time as a reasonable adjustment

An important area where test speededness can have significant consequences is in the use of extra time awarded to test takers as an assessment adjustment. The Equality Act 2010 places an obligation on schools, colleges and awarding organisations to ensure that individuals are not disadvantaged in assessments due to any protected characteristics. One part of this is that reasonable adjustments should be made available to address any substantial disadvantage that a disabled person would have in undertaking an assessment.

The act does not specify the adjustments that should be made available as this is the responsibility of the school, college or awarding organisation. A variety of adjustments are available in practice, but the adjustment most commonly granted by awarding organisations is 25% extra time to complete written tests. Reasonable adjustments are available to address a variety of needs, and extra time is often made available to individuals with problems in speed of processing language and understanding questions, response planning, and producing written responses[9].

Similar legal contexts exist in many other countries. Lovett (2020) discusses issues around extra time in time-limited assessments in the US context in detail. He describes the conventional view in validity theory that for high validity a test should have standardised testing conditions (including time limits) for all individuals, which allows test users to draw valid interpretations of the test outcomes. Differences in testing conditions can introduce a threat to validity. However, even aside from the legal requirements, if an individual has a disability that slows down their speed of responding but they sit the test under the same time limits as others, this can threaten the valid interpretation of their test outcomes if speed is not part of the test construct. If speed is an intended part of the test construct, and an individual’s deficit affects the very thing the test is intended to measure (for example that a task can be completed within a fixed time, without exception), then it is likely that the provision of extra time would not be appropriate. Other adjustments may be possible in these circumstances.

The provision of extra time as a reasonable adjustment for assessments of knowledge, skills and understanding is intended to equalise opportunity to demonstrate achievement across those test takers with and without extra time. To do so, those test takers without extra time should be able to maximise their performance within the standard time limits, and it is only those test takers with specific needs who receive extra time whose performance is limited[10]. This has been termed the ‘interaction hypothesis’ (see Sireci, Scarpati and Li, 2005) because it is only those receiving extra time who should show improved outcomes with more time. An alternative perspective on the same idea is the ‘maximum potential thesis’ (Zuriff, 2010), in which after the application of the extra time adjustment, both groups should have equal potential to show their maximum performance. Note that this is not about equalising outcomes, but about equalising opportunity. Three conditions need to be met for this thesis to hold. First, the test should not be speeded for those taking the test under standard time limits. Second, the right individuals, and only those

[9] Extra time may also be granted for reasons other than disability, such as for temporary illness or injury. The same theoretical considerations apply in these cases, but here the discussion concentrates on its provision for disabled people, reflecting the focus in the research literature.

[10] In theory, extra time could also be used to equalise the speededness of a test that is speeded for those without extra time. This would require precise measurement of test speededness across test takers, and precise allocation of extra time to individuals. At the individual level, each person might need a bespoke test duration. Even at a group level – for example comparing those with and those without extra time – it would be difficult to achieve equality of speededness in practice. This approach is not mentioned in discussion of extra time in the research literature.

individuals, receive extra time. Third, the extra time awarded should be sufficient to allow those receiving extra time to show their best performance.

A consistent finding in the research literature is that in many tests, most test takers, regardless of disabilities, appear to benefit from an extra time allowance by scoring more highly, as demonstrated across 3 separate systematic reviews (Cahan, Nirel, & Alkoby, 2016; Lovett, 2010; Sireci, Scarpati, & Li, 2005) as well as a review by Ofqual (Cuff, Keys, Churchward et al, 2025). Broadly, those individuals who would normally receive extra time benefit more from additional time than those who do not, but many test takers not normally receiving extra time do improve their scores[11]. This suggests that many of the tests that were included in these reviews were speeded. However, the reviews show variable effects of extra time, supporting the idea that some tests are more speeded than others. There will be tests with appropriate, or even generous, time limits, where extra time does not improve the performance of test takers who do not normally receive extra time. It is unclear where on a scale of speededness exams for GCSEs, AS and A levels sit. But if they are somewhat speeded, many of those without extra time might perform better with more time, and therefore extra time could give those individuals receiving it an advantage over those without.

One measure that can be used to determine whether test conditions including adjustments are fair across all test takers, and that test scores have the same meaning regardless of test adjustments, is to look at the predictive validity of tests for those with and without extra time. That is, how well the test scores predict future outcomes on other related tests or real-world tasks. Lovett (2020) reviews evidence of how well university admissions tests predict later performance (grades, or completion rates, for the first undergraduate year or full degree) on law or medical degrees. The evidence suggests that outcomes were worse for students who received extra time than for those without extra time, in groups matched for admission test scores. This would suggest that the admission test scores may not be truly equivalent across groups, with those receiving extra time achieving relatively higher scores than equal-ability individuals without extra time. However, limitations in the data such as a lack of control for other factors affecting degree completion, and a lack of information as to whether extra time was available for exactly the same students on the degree level tests as for the admissions tests, make any conclusions tentative.

In England, the Independent Commission on Examination Malpractice (ICEM) report (Joint Council for Qualifications, 2019) considered the issue of the award and use of 25% extra time as an access arrangement because there had been some concerns about the reported increasing numbers of students receiving extra time, and whether

[11] This has been termed the 'differential boost' hypothesis.

there might be some abuse of the system. They suggested that some analysis of qualification outcomes for students with and without extra time should be carried out, looking at groups of equal ability. At the time of writing in 2025, this has not been carried out. Precisely matching groups with and without extra time on ability may be difficult since some valid external measure of ability in the subjects analysed based on untimed testing would be required, otherwise ‘ability’ would be confounded by speed of working. Therefore, it may be that looking at predictive validity of outcomes for equally-achieving groups could be worthwhile. The ICEM report also quotes unpublished research from Cambridge Assessment from 2016 which showed that while students with learning difficulties demonstrated significantly higher test scores with extended time, so too did students with no learning difficulties, in agreement with the literature reviews quoted earlier. It is not clear which tests were analysed or how.

As noted earlier, a long-standing view in validity theory posits that fairness is achieved through standardised testing conditions, except where some individuals experience a clear deficit in their ability to access the test and an adjustment (referred to as a ‘test accommodation’ in the research literature) is appropriate. Ideally, the test should be designed such that as many individuals as possible can access it under standard testing conditions, minimising the need for adjustments (the principle of universal test design mentioned earlier). This is also reflected in Ofqual’s General Conditions, which require awarding organisations to design their assessments such that the need for reasonable adjustments is minimised (General Condition E4.2d). More recently there has been some move towards the idea of more personalised assessment (see, for example, Sireci, 2020), although this is perhaps difficult to imagine implementing in the high stakes testing context in England. A recent book by Nisbett and Shaw (2020) took an in-depth look at issues of fairness in assessments from a largely English (or at least anglophone) perspective. Many of the issues of fairness they discuss revolve around equality of test outcomes for individuals and groups with identified or protected characteristics through well-designed and thought-through assessments.

The authors do not directly address the provision of extra time in the context of possible test speededness. They do, however, mention the difficulty of ‘levelling the playing field’ when it is unknown how a candidate would perform if they did not have a need requiring an adjustment (in other words under no time constraints). They do not reach any clear conclusions on how test adjustments should be decided. The thinking that has gone into this book and the issues that it raises illustrates how difficult it is to decide how far to go in making adjustments to test design or test conditions to ensure fairness.

A review of methods of measuring test speededness

So far general issues of time in assessments and when speed of working may be a valid part of a test have been discussed. To progress our thinking, particularly around exams in GCSEs, AS and A levels, some consideration of methods of estimating or directly measuring test speededness is needed. Being aware of methods to measure test speededness will help inform investigation of tests in England, since a key unknown remains whether certain tests in England are speeded, and if so, to what extent.

When talking about tests being speeded, it is important to consider exactly what is meant by this. When a method of measuring speededness is applied to a test, it is really measuring the speededness of that set of test scores, not the test itself. If a different set of individuals sat the test, the results of an analysis of speededness may be very different. Similarly, if a different instance of the same test (say, a paper from another year or a different set of items from an item bank) was sat instead, again, conclusions about test speededness may differ. This means that unless the analysis of speededness is carried out across multiple test versions and cohorts, care is needed when extrapolating from such an analysis to draw conclusions about the speededness of a test in general. Further care in using and interpreting speededness measures is also required. First, as noted earlier there is no really clear dividing line between a test being speeded or not, any threshold is likely to be arbitrary and hard to defend. Second, reducing speededness down to (sometimes) a single number can also downplay or ignore a wide range of individual experiences, with an average not being representative of the experience of those most affected by time limits.

Before exploring the different methods of measuring speededness in the literature, it is worth noting upfront that a great deal of the research into test speededness has been carried out in the US, and this has tended to focus on tests composed of single mark, usually multiple-choice type questions. Little research on how to measure test speededness has been carried out on tests with constructed response items. As Lu and Sireci (2007) note:

> “Unfortunately, literature on evaluating speededness for constructed-response items is scarce. While a random guessing strategy is commonly used under speeded conditions to answer multiple-choice items, it is not really relevant with respect to constructed response items. Rather, with insufficient testing time, examinees usually leave constructed-response blank or provide answers that are much shorter than they would be if there were sufficient time. Blanks and shorter

> responses could be signs of speededness or of lack of motivation, particularly on low-stakes tests.” (page 31)

This observation is true today, as most research since 2007 has focused on the integration of item response time information in measures of speededness for multiple-choice tests, which computer-based testing has allowed.

However, there is still value in understanding how test speededness has been measured, albeit mostly in multiple-choice tests. There are 2 main approaches. A more experimental approach requires a test to be administered more than once, under different timing conditions. This experimental approach can be more difficult to implement for live test administrations, as described below. The alternative approach is to estimate speededness from a single test administration, often using statistical models, and many approaches have been developed that take this approach. As will become clear, this approach requires some assumptions to be made, which do not always stand up fully to scrutiny.

This is not intended to be a thorough review of all the different models, but to give a flavour of the main approaches. Bridgeman (2020) and Jurich (2020) in the NCME book, as well as a recent paper by Cintron (2021) review the history of methods used to measure test speededness in some detail. While these tend to be quite US-centric in the research they describe and research does take place elsewhere, in Europe for example, these sources provide a good overview of the main approaches that have been developed. Bridgeman identifies several phases, with earlier research starting in the 1940s focusing on item completion statistics to identify speededness. Then, from the 1970s concerns about differential speededness on tests for test takers with different characteristics (such as sex and ethnicity) arose, and some experimental studies were carried out together with the development of more complex statistical approaches. Finally, with the introduction of computer-based testing (CBT), including computer adaptive testing (CAT), analyses that included item response time became possible, including the detection of rapid guessing behaviour on multiple-choice items.

Multiple test administration research

At their simplest, multiple test administrations allow the difference between scores on a timed test under its normal time limit to be compared to scores with either more generous timing or unlimited time. This approach should give a direct measure of the effect of the time limit on scores, assuming all other variables such as test-taker motivation are kept constant. Kane (2020) terms the difference between scores in these 2 conditions when speed of working is not part of the intended construct ‘time-limit errors’ (TLEs). For a speeded test, TLEs would be non-zero, and across different tests TLEs for each test taker would be correlated to a degree since they represent an underlying 'speed of working' measure for that individual. These TLEs therefore introduce systematic, not random, error. In other words, they introduce construct-irrelevant variance rather than just measurement imprecision to the test.

Comparing performance on a test under different time conditions gives a direct measure of time-limit errors, and requires very few assumptions to be made to calculate test speededness. However, it can involve some practical difficulties. One approach is for 2 independent groups of test takers to take the same test under different timing conditions. Either the same test is taken with different time limits, or different length tests are taken under the same time limits. Such a multi-group approach may require precise matching of the (relevant) demographic characteristics, as well as abilities, of the 2 groups of test takers, which may prove difficult to achieve in practice. Rather than trying to match groups, individuals can be assigned randomly to groups, but these groups will need to be larger to ensure the group characteristics are sufficiently well balanced.

An alternative approach is to use parallel test forms, which allows effectively the same test to be administered more than once to a single, smaller, group of test takers. Ensuring equivalence of the parallel test forms is vital, as is ensuring that test-taker motivation is maintained equally across the multiple administrations. Because this approach is likely to appear obviously experimental to the test takers – why else would you as a test taker be sitting multiple versions of the same test? – there may be a greater risk of low motivation that could disrupt accurate measures of speededness.

Cronbach and Warrington (1951) suggested a theoretical design combining both approaches – 2 independent groups each taking 2 parallel versions of a test, one test with unlimited time and one under time-limit conditions, and with the test versions crossed across conditions for the 2 groups. This design meant that exact test equivalence would be less critical. A coefficient could be derived from the correlation between scores under the 2 different time conditions for the 2 test forms, which would represent the degree of test speededness. This approach is numerically quite simple and requires no assumptions about how speededness manifests in the test but suffers from its obviously experimental nature, risking low test-taker motivation. It does not appear as though this design was ever used in practice.

From the 1970s, concern around sub-group differences in the impact of test speededness were raised. Evans and Reilly (1972) used one of the first experimental approaches by comparing a reduced-length version of the Law School Admissions Test (LSAT) they produced with the standard form under the same time conditions to determine whether black students were differentially affected in the more time-limited condition. While performance improved overall with the shorter test form, they found no convincing evidence of group differences. Wild, Durso and Rubin (1982) took the alternative approach of administering a fixed, unscored section of the

Graduate Record Examination (GRE), under different time conditions at different test centres. They were able to use performance on an unmodified live section of the same test sitting as a covariate to control for any differences across the groups taking each test form. They also observed better performance for the more generously timed version of the modified test section, but there were no differential effects of sex or ethnic group, suggesting all sub-groups gained equally from more time.

From around the turn of this century, the focus on sub-group effects shifted to whether disabled test takers were being advantaged by extended time limits (extra time) because those sitting the test with standard time were experiencing the tests as speeded. A motivation around this time was that there had been some concern in the US that extra time could potentially be abused through parents obtaining paid-for diagnoses for conditions to access extra time.

Various studies utilising different length sections of tests taken under fixed timing conditions were carried out to look at the overall speededness of the tests under standard time. Bridgeman, Trapani and Curley (2004) looked at performance on a pilot section of the SAT test not used for grading by creating versions of this section with different numbers of questions by deleting carefully selected questions to maintain equivalent difficulty across the test versions. They found that scores were lower on all items for the longer test forms. Similar findings were found for computer administration of the ACT by Li, Yi and Harris (2016) and for the USMLE by Harik, Clauser, Grobovsky et al (2018) as described earlier. The overall evidence that accrued was that tests were often speeded to a degree for many test takers, including those not receiving extra time.

Multiple test administration methods should achieve relatively uncorrupted measures of speededness, without requiring assumptions about how test takers use time in the tests, or the sequence they completed items. However, apart from cases where a test includes sections that are not used for grading (typically these are for trialling new items), which can therefore be experimentally manipulated, there are logistical challenges to running such studies. In England, because of the security concerns around new papers, such research would be hard to carry out. New, bespoke papers might need to be developed for purely experimental studies, with substantial associated costs. Studies using existing older papers might be possible, except for the high likelihood of the paper having been seen by many potential test takers in mock exams or general test preparation in class. Because of the general challenges of running multiple test administrations, more effort in the US has gone into analysing data from single test administrations.

Single test administration measures of speededness based on item completion

The first methods to analyse test speededness using single test administrations relied upon measures of omitted items, mainly items at the end of tests classified as ‘not reached’ due to time constraints. Later approaches also considered patterns of response correctness to look for signs of guessing (in other words, performance at chance level on multiple-choice items). Methods that use statistical models, including those including item timing information from CBTs, are described in the next section.

As noted earlier, Gulliksen (1950) first formalized the distinction between speed and power tests. Gulliksen was also one of the early researchers who looked at quantifying speededness using statistical analysis of single test administrations. The Gulliksen metric was a ratio comparing the standard deviation[12] across all test takers of the number of items not reached at the end of the test to the standard deviation of the wrong items (incorrectly answered plus not reached). This analysis assumed that no items were omitted earlier in the test due to time. If a test was not speeded, for all test takers the not-reached items would be zero, giving zero standard deviation, and thus the ratio would be zero. Where the test was speeded, there would be a range of not-reached item counts, giving a non-zero standard deviation. It was expected that this standard deviation would be smaller than the standard deviation of wrong items, giving a ratio less than 1. However, it was found that in some analyses this ratio could exceed 1, and so Gulliksen suggested that other ratios between different item classifications should be considered, but these proved hard to interpret and draw firm conclusions from across all degrees of test speededness.

Stafford (1971) aimed to solve the problems with Gulliksen’s approach by suggesting a simple sum calculation, where the ratio of the sum of not reached items for all test takers and the sum of errors (wrong answers including not reached items) gave a ‘speededness quotient’. This was simply the proportion of wrong responses that arose because the item was not reached. Because items during the earlier stages of a test may be omitted, Rindler (1979) also noted that it would be necessary to make a further distinction between omitted items (unattempted items surrounded by attempted items) and not-reached items at the end of the test.

Jurich (2020) notes that these statistical approaches did not appear to become widespread in use. Rather it was the simpler metric proposed by Swineford (1956, 1974), based on percentages of items reached, that appeared to be widely applied, despite the arbitrary nature of these values as Swineford himself conceded. As a rule

[12] A measure of the spread of a set of numbers.

of thumb, Swineford suggested that all test takers should reach the 75% point in a test, and 80% of the test takers should reach the end. However, apart from the arbitrary cut-offs, these metrics assumed that test takers always attempted items in the order they appeared on the test (no skipping, or jumping back) and also neglected any more subtle effects of time limits, such as hurrying on attempted items. However, such rules of thumb appeared to have had more traction than the calculation-based methods above, and indeed these guidelines appear to still be captured to a degree in the current Standards for Educational and Psychological Testing (AERA, APA, NCME, 2014). The comment for standard 4.14 quoted earlier mentions checking the proportion of test takers who reach the end of the test and also checking item completion rates.

One further, more experimental, approach to single-test administration measures is to administer a test with a time limit, but then to allow additional time after this and determine how many additional marks are achieved. Oppler, Davies, Lyons, Nathanson, & Chen (2004) used this method with the Medical College Admission Test (MCAT), getting test takers to change the colour of the pencil they used to give responses at the end of the standard time. They found that test scores increased with more time, because more questions were attempted, but also through checking and amending previous answers (the MCAT is a multiple-choice test). There was also evidence of variation in the benefit of additional time across ability level. One concern that arises from this design is that it is likely to be more effective at determining the effect of additional time if test takers are unaware that this time will be given at the end, since they will then apply greater effort to the standard-time part of the test. This requires deception (or at least a lack of transparency). This is what Oppler et al did. Deceiving participants in experiments is perhaps less acceptable today than it was then. Such an approach also feels less appropriate in high-stakes conditions, and thus the confounding effect of lower motivation in low-stakes conditions may undermine the method.

As greater information such as item response time became available from computer-based testing (CBT) from the 1990s, more complex statistical models were developed. These are discussed next.

Model-based speededness measures including those that use response time

Many tests worldwide, but particularly in the US, are underpinned by item-response theory (IRT) models (Hambleton, Swaminathan, and Rogers, 1991). These use data from piloting of the test or test items, sometimes supplemented by ongoing live test administration data, to estimate item difficulty from the pattern of marks achieved by test takers. This data can be used to construct tests of appropriate difficulty, or set up adaptive tests where the items presented have difficulties appropriately matched to the estimated ability of the test taker. Nearly all of the research and modelling looking at test speededness has been applied to cases where the tests have single-mark items, and in nearly all instances these are multiple-choice items. However, variants of IRT models exist that account for multiple-mark items, and in many cases these might be simply swapped into the speededness models to allow them to work with more varied test items. The fact that they are rarely discussed or trialled in the literature emphasises the different testing context in the US.

The most commonly used IRT models assume unidimensionality – that the test is measuring a single dimension, that of ability in the targeted construct. Any test speededness violates the assumption of unidimensionality by introducing another dimension, speed of working, that determines item responses in addition to ability. The construct-irrelevant variance speed of working introduces can distort IRT item difficulty parameters where items are skipped or hurried over, making the items appear more difficult than they really are. This means that to interpret IRT item difficulty data, the assumption that tests are not speeded has to be made (Hambleton and Swaminathan, 1985). This problem motivated methods to determine where items were speeded and to remove any effect of speededness from the estimates of item difficulty.

While the early statistical attempts to model speededness relied on omitted items at the end of tests, it became clear during the 1990s following the availability of item response time data from computer administration of tests that speededness did not manifest only in this way. Schnipke (1995) noted that response time data showed that test takers were sometimes answering questions very rapidly, too fast to represent full consideration of the item. In fact, there was evidence of a bimodal response-time distribution for many items. One part of the response-time distribution, seen as a broad normal (or gaussian) distribution represented attempts to apply knowledge and skills to answer the item – what Schnipke termed solution-oriented behaviour. However, there were also many responses that had only taken a few seconds – far less than the time needed to fully consider the item. This was evidence of rapid guessing behaviour, where a random selection from the multiple-choice response options was made.

Schnipke and Scrams (1997) modelled response time data using 2 (mathematically, log-normal) distributions. Because in general the amount of rapid-guessing behaviour increased as the test progressed, they modelled a single point for each test taker where they switched from solution-oriented to rapid-guessing behaviour. They found that this 2-strategy model fitted the data much better than a single-strategy model, especially in the later part of the test where guesses were more prevalent. However, it is worth noting that Schnipke (1995) showed evidence of a non-monotonic increase in rapid guessing throughout the test, with some items, including some earlier in the test, showing more evidence of rapid guessing than nearby items. Therefore, test takers were not simply working through the test start to finish exhibiting full solution-oriented behaviour until time was almost up and then guessing the last set of items. Their behaviour was a little more complex than this. A later model by Wise and DeMars (2006) proposed a similar approach to Schnipke and Scrams, but allowed behaviour to switch repeatedly between solution-oriented and rapid-guessing behaviours, reflecting this more nuanced view of test-taker behaviour.

Where item response-time data was not available, response accuracy itself could also be used to identify signs of rapid guessing for multiple-choice items, since accuracy should fall to chance levels. Because this would also be true of items that were too difficult to answer, this approach was restricted to the analysis of patterns at the end of tests. The HYBRID model of Yamamoto (1995) was based upon estimating the point in the test at which accuracy consistently fell to chance levels, with test takers switching from solution behaviour to rapid guessing. The distribution of these switching points over the population of test takers could be analysed and used as a measure of overall test speededness.

The HYBRID model itself comprised a standard IRT model to estimate item difficulty and test taker ability parameters based on data up to the point at which the switch from solution behaviour to random guessing occurred, and then used a fixed random guessing factor after that. The switch in strategy could only occur at one point for each test taker, and random guessing could only occur in items after the switch point. The optimum model fit gave the individual switch points for each test taker. Because there is only one switch point for each test taker the HYBRID model is limited in applicability to tests where guessing occurs mostly only at the end. As already noted, this may be a questionable assumption in many cases. Later models allowed a gradual shift from solution to guessing behaviour rather than a fixed switch point, but were otherwise similar.

Another approach by Bolt, Cohen and Wollack (2002) separated candidates into 2 separate speeded and non-speeded groups based on evidence of random guessing behaviour from their response accuracy. Item difficulty parameters for the speeded items were estimated using a Rasch model separately for the 2 groups so that those of the non-speeded group were not distorted by speededness. The number of test takers classified into the 2 groups could be used as a measure of test speededness. Demographic differences in the group membership were also used to check for any differential impact of time limits. Further development of this approach led to the multi-class mixture Rasch model (Mroch, Bolt and Wollach, 2005) where rather than only having 2 groups, there were multiple groups based on the item at which speededness was first detected (much like the HYBRID model). The main contrast with the HYBRID model is that speededness leads to random guessing in the HYBRID model, whereas it leads to increased item difficulty in the Bolt et al and

Mroch et al models, which accommodates the effect of hurried but not random responses.

More recent models have also used response accuracy data as a way to identify speededness effects. Shao, Li and Cheng (2016) for example used change point analysis to determine where test takers first experienced a test as speeded. IRT ability estimates were calculated from near the start of the test and then updated by incrementally adding more items to the calculation, and an item at which a significant drop in the ability estimate occurs (because the item was answered incorrectly when it would be expected that the test taker could answer it correctly) indicated the effect of speededness. Similar to other models, these speeded individuals, or item responses, could be excluded from item difficulty estimates.

Later approaches included item response time as well as response accuracy in the model fits. Van der Linden (2007, 2011) combined traditional IRT person ability estimates with individual speed estimates in a hierarchical model. Two separate models were used to estimate person and item parameters based on response accuracy and, separately, response time, then these separate models were combined in the higher-level model that determined the relationship between them. The speed model calculated a single speed parameter for each person, which was assumed to be constant across a test – as if each test taker decides what the appropriate pace is for them across the test, but that time per item varies with the difficulty of the item (more time will be spent on harder items). This model, and variants of it, have been widely researched and various analyses have been carried out looking at the relationship between item difficulty, time spent on items, likelihood of a correct answer, and test-taker ability.

Models using both response time and response accuracy are often designed to improve the accuracy of ability estimates in tests where partial or rapid guessing occurs. This may be because of either low motivation in low-stakes conditions (for example, Wise and Kuhfeld, 2021) where scores tend to be lowered, leading to underestimated ability parameters from the model, or to correct ability estimates when high-stakes tests with no penalty for guessing are speeded (for example, Feinberg, Jurich and Wise, 2021). The timing information obtained from the models can also facilitate test development as summarized by van der Linden (2011, 2017). Test time limits can be set based on the modelled time required for individual items. Although of course there is variation in the time individual test takers will require to answer a set of questions, the model data can help to design tests with an estimate of how many test takers may experience the test as speeded, and to what extent. Similarly, alternate test forms can be created that are matched on both difficulty and estimated time required to respond.

The van der Linden (2011) approach assumed a fixed relationship between the speed an individual test taker worked at and the time required for all items (a faster test taker would have a reduced time per item for all items compared to a slower test taker). Later developments (for example, Becker, Weirich, Goldhammer and Debeer, 2023) have extended this model to allow variable ‘speed sensitivity’ for items. This means that that time required to complete some items will vary less than other items across test takers who work at different speeds. Research using such hierarchical IRT models continues today, but because these models are not directly applicable to paper-based tests with constructed-response items the full range of studies using them is not described here. For more complete details of the significant number of models that have been developed, the reader is referred to Lu and Sireci (2007), Jurich (2020) and Cintron (2021).

Speededness of tests with constructed-response items

Most of the research and the models that have been developed, both in the US and elsewhere, have been based upon tests composed of single mark, selected-response test items, such as multiple-choice questions. Much less effort has been expended looking at speededness of tests with constructed response items, despite such items occurring in some graduate admissions tests (such as the GRE mentioned earlier) and various credentialling tests in the US and a variety of tests worldwide. From the perspective of GCSEs, AS and A levels in England, it is important to understand the effect of time limits on these item types.

Margolis, von Davier and Clauser (2020) report on some studies that investigated time limit effects in exam essay writing. Their review shows variable evidence, such that in some circumstances (for example, Biola, 1982) more time can lead to higher scores, while in others (for example Caudery, 1990) no benefit was apparent. The reviewed studies appear to be generally low-stakes in nature for the test takers. This makes interpreting any of these findings difficult, since stakes are likely to impact the effort made in writing an essay response (see, for example, Robin and Zhao, 2014). Variation in the findings probably also reflect how generous the time limits were relative to the time needed to write a credit-worthy response for each essay. Broadly, this research on low-stakes tests showed that improvements in scores were often seen with more available time, and there were no differential effects across groups or ability level.

One study (Klein, 1981), which may have involved more motivated test takers, used an optional essay section on a California Bar Exam with 2 essays to be completed under different time conditions, using a counterbalanced design with both essay questions occurring in both time conditions. Because test takers were told that this section might be used to make pass decisions if they failed on the rest of the test, these essays may have had relatively high stakes for them. The study found that higher scores were obtained for the essay with more generous time allowed, and that this was true for both essay questions.

Margolis, von Davier and Clauser (2020) also describe research looking at computer-based case simulations in the USMLE. Although not a traditional academic test requiring an essay response, these simulations require a scored constructed response as the outcome measure, although the time allowed per item also includes navigating the case simulation and drawing conclusions, not just writing a response, so this is not a direct analogy of extended-response questions. When test takers tackled versions of the simulations with 15, 20 or 25-minute durations, Margolis et al noted a general reduction in performance with less available time. However, the largest performance decrements were seen between 15 and 20 minutes, suggesting that the standard 25-minute time was fairly generous and unspeeded for most. They looked at the proficiency of the test takers, and found variable effects across different simulations. On most simulations, all test takers were affected similarly, but on some, the most able were barely affected by reduced time limits, while time-limit effects were most severe for those of moderate ability. These findings highlight that ability-specific effects of time arise across different test types, not just those using multiple-choice questions.

Methods of investigating speededness in tests composed of constructed-response items have been mostly experimental, requiring test (or item) administration under different time limits, rather than using statistical models based on single test administrations. This is understandable since the effect of time limits may be quite subtle and variable across items on the test. If time is limited, missing or extremely brief responses may indicate the effect of time, but of course may also indicate the effect of item difficulty if the test taker could not answer the item well.

Finally, for longer constructed responses such as essays, time is also needed to plan a response, not just write it. More generous time may not necessarily increase the length of responses, but may allow test takers to formulate more credit-worthy and coherent responses without increasing response length. The effect of limiting time for planning is difficult to measure without explicitly separating planning time from writing time. The only study in which additional planning (but not writing) time was provided (Hale, 1992) showed no improvement in scores, but this was in a low-stakes context, and so not an ideal test of the use of planning time.

Limitations in measures of speededness and the effect of test-wiseness

As Bridgeman (2020) puts it, the gold standard of speededness research is to run multiple timing conditions for a test with independent random groups assigned to each condition, rather than using more complex statistical analysis or modelling techniques on single test administrations. However, as noted, experimental studies can be difficult to carry out in practice, whereas statistical approaches can in theory be applied to any test that has been sat by a reasonable number of test takers.

The analysis of single test administrations requires assumptions to be made regarding the behaviour of test takers. Most of these assumptions turn out not to always hold, although perhaps some may have been more reasonable in the middle of last century when some of the methods were devised and test takers may indeed have been more naïve, and certainly not so thoroughly coached in how to take tests. Both analysis of omission rates and rapid guessing on items towards the end of tests required the assumption that test takers progressed start to finish through the test items at their own preferred pace until time ran out, leaving unattempted or guessed items at the end of the test.

This ignored 2 factors. First, that omission rates are likely to be misleading because in multiple-choice tests where there is no penalty for incorrect answers[13] it would be highly naïve behaviour when time is running out not to complete the remaining items by rapidly guessing them. To complicate matters further, the impact of time limits does not only affect items at the end of a test. While Schnipke (1995) showed that rapid guessing was more common at the end, it could also occur throughout a test, probably related to item difficulty where an item looked like it would take an unjustifiably long time to answer, if it could be answered at all. Similarly, if items were to be omitted, this might also occur for difficult items earlier in the test.

Second, the idea of a binary distinction between solution-oriented behaviour where the question is given full consideration, and rapid-guessing (or omission) where there is no consideration of the item is simplistic. The introduction of CBT during the 1990s provided rich response-time and item-sequencing data which showed that test takers behaved in a complex fashion. They skipped items to return to later, or varied the time spent on each item (for example the work of Bridgeman et al, 2004, Harik et al, 2018, described earlier) by adjusting their location on the speed/accuracy continuum as available time allowed. This included making partial (in other words, informed) guesses on multiple choice questions, where some consideration of the question occurs, allowing some response options to be eliminated, which would lead to above-chance performance. All of these factors mean that analysis of omission rates or rapid guessing only captures some of the effect of limited time on item responses, and so may represent an underestimate of test speededness, or at least a lower bound on the estimate. Of course, if a lot of items are omitted due to excessive difficulty, omission rates may over-estimate speededness.

[13] Some tests do penalise incorrect responses and in these cases no response would be preferable to a completely random guess.

All of the above observations are manifestations of the amount of test-taking nous, known as test wiseness, that individuals bring to the testing experience, which has increased over the years. In the early days of speededness research, while not all test takers were necessarily naïve, strategies and techniques to maximise scores on tests were not widely taught. Increasing pressure to perform and competition has led to a greater focus on formal preparation and strategy use during tests. Today, any suggestion that test takers approach a test that has high stakes for them with no preparation and thought as to how to use their time cannot stand up to scrutiny. In fact, 'teaching to the test' has become a concern over recent decades (for example, Koretz, Linn, Dunbar and Shepard, 1991; Cizek, 2005), with the fear that there is too much focus on preparation for tests which distorts the taught curriculum and can limit actual knowledge and skill acquisition by students.

Use of test results for accountability (for example teacher, school or educational system performance) has been one driver for a focus on test preparation in teaching. In England the introduction of school performance tables from 1992 (see, for example, Leckie and Goldstein, 2017) would have focussed the minds of teachers on test preparation. In high-stakes performance-table qualifications in England, class teaching time will be dedicated to preparation for the test. As well as practising answering specific questions, this can include sophisticated teaching on how to monitor elapsed time against progress through the test so that pace can be adjusted if necessary, and how to apportion time in different ways to different questions, particularly where the questions have different maximum marks.

The opportunity to use strategy in taking a test does vary depending on the test design, but in the context of exams in GCSEs, AS and A levels, most allow the freedom to answer questions in any order. Preparation may focus on the optimum order in which to answer questions based on those questions the individual is most likely to be able to answer most efficiently, judging likely marks obtained against the time taken. The aim of all of this test preparation is to use the available time in a way that maximises the marks each individual is able to achieve.

However, even for tests that are not used for school accountability, while they may not receive direct teaching and instruction, if the stakes are high for individuals they will often independently prepare themselves by taking practice tests allowing them to hone their technique. Therefore, today's test takers start a test with a good idea of the format of the assessment, the type of questions that will be asked and how time limits may affect their performance on the test.

Overall, test-wiseness suggests that the binary distinctions that many measures of speededness require do not hold. In most time-limit tests there may be no such thing as 'full consideration' of an item, since the test taker may have chosen to take longer on the item if more time was available. They may have spent more time planning and thinking, more time writing (if it was constructed-response) or more time checking their answer. Therefore, they metaphorically choose a position somewhere along the speed/accuracy continuum, balancing time per item with response accuracy (or completeness).

Adapting the response time allocated to items may often mean that a test will appear unspeeded from a single test administration since there may be no detectable changes in speed, obvious rapid guessing or full item omission that can signal speededness. It seems probable that even on a test that appears to be unspeeded under its standard time conditions, potentially all items may be speeded to a degree, in that the test takers may have chosen to use more time on the items if it were available.

This is not to say that the early researchers were naïve or wrong, just that how to account for more complex test taker behaviour was not clear, and there was a process of model development and extensions that had to occur. Each new development would account for more factors, which should in theory improve the fit of the model to the data. Even the early hierarchical IRT models (for example van der Linden, 2007) assumed a fixed person speed parameter, indicating their individual tendency to work at a particular speed. This then excludes the possibility of each person changing their speed of working as the test progressed, in an individual way. Later models contain more parameters to model more individual factors, but a downside is that because the models have become increasingly complex and mathematical, interpretation of their outputs has become more difficult, and they may be lacking transparency (Jurich, 2020).

A final caveat noted by, for example Jurich, 2020, is that much of the research that has been carried out has been on low-stakes tests, in which test-taker motivation tends to be lower. Because the test outcomes may have little consequence for the test takers, some may not be motivated to apply their full effort to the test, giving only partial consideration to items, adopting rapid random guessing just to ‘finish’ the test, or omitting larger numbers of questions. Therefore, speededness measures may be affected and not reflect the speededness a motivated test taker would experience. Ideally, to know whether time limits are truly impacting performance, testing should take place under high-stakes conditions so that test takers are motivated. The more convincing studies are those which have looked at high-stakes testing, often admission tests for university, such as the SAT, American College Test (ACT) and Graduate Record Examination (GRE), where outcomes are important to the individuals.

As noted already, all of these considerations around use of time and measuring speededness become more difficult when considering tests with constructed response items, such as in the English context, to which we now turn.

Approaches to measuring speededness in the testing context in England

As will be clear by now, tests composed of constructed-response items, particularly paper-based tests, have not been a major focus of the research into speededness. To evaluate the speededness of paper-based constructed-response tests, experimental approaches that involve multiple test administrations with different timing conditions might be needed. The alternative experimental approach of switching colour pens (or pencils) at the end of an initial time limit and then continuing the test for an additional period and comparing marks could also be used.

The practical difficulties of either approach have been noted. The research reviewed earlier does include several examples of using pilot sections of larger tests, often used for trialling new items, which can be co-opted for experimental work. Unfortunately, this type of pre-testing is not used for GCSEs, AS and A levels in England. Live GCSE, AS and A level exam papers (and a variety of other qualifications) are sat once, on a single day, then retired. Because of this any advance use in experimental contexts would present major security concerns, and later usage would risk pre-knowledge on the part of test takers, since these papers (and individual questions) are often used for class tests and mock exams. This means that bespoke papers might be required for use in experimental situations. The need to ensure comparability so that any conclusions drawn were valid would mean that these experimental papers would have to go through the full quality assurance process that exam boards use for live papers.

Therefore, do any of the existing single administration methods for measuring speededness appear applicable to GCSEs, AS and A levels? It helps to split tests into those that contain a substantial number of lower-tariff items and those that have a smaller number of higher tariff extended-response items. Maths and to a lesser extent science papers, in GCSEs particularly, exemplify the first category, with often more than 25 separately scored questions. English language, English literature and history (as well as a variety of other essay-based subjects at both GCSE and A level) exemplify the latter design, with sometimes as few as 2 questions to be answered (often from a choice of optional questions). Of course, assessments in subjects represent a continuum, and there are subjects such as psychology or geography that are intermediate in design with both extended-response and short-answer questions on their papers.

Tests with many lower-tariff items

In a subject such as GCSE mathematics, with somewhere around 30 questions on an exam paper, of which most have 1 to 3 marks available, there is a strong tendency to answer sequentially (Spalding, 2010). The mental 'cost' of jumping around the items on the paper and keeping track of them may be quite high for individuals, especially given how linked sub-questions relate to a similar context. This is not to say that strategic skipping over of questions (perhaps for later consideration if time allows) will not be common. In GCSE science papers, which have a similar number of questions, there may be a slightly greater tendency to sequence groups of questions (or sub-questions within a whole question) because of the way they are clustered within topic areas. For example, there may be 4 or 5 partially or fully linked questions on mechanics. However, a sequential approach with judicious skipping may still be most likely.

Because of all the noted effects such as partial attempts at these short-answer questions, any incomplete or truncated responses may be due to limited knowledge as much as to limited time, and so the only real indication of test speededness may be based on not-attempted items. The least ambiguous indication that test time limits may have been a factor is where there is a continuous run of items towards the end of a test that have not been attempted, classified as not-reached items. Because in many tests item difficulty is intended to broadly increase from start to finish, item difficulty may also have contributed to the last items not being attempted. However, if some control for test-taker ability and item difficulty is included in an analysis, these not-reached items at the end of a test may indicate a lower bound for the effects of test speededness, as noted earlier. There may be many other subtle effects of time limits in attempted questions, and so speededness may be higher than the not-reached item analysis suggests.

Not-attempted items earlier in the test which are bracketed by attempted items may occur because the item is too difficult as well as because of test time limits. Even allowing for modelling the effect of test-taker ability it may be difficult to distinguish between an item omitted purely based on difficulty (it would not have been attempted even with unlimited time), and one omitted because limited time meant that a strategic decision to skip it was made, perhaps with the intention of returning to it later on if time allowed. This means that the significance of these earlier omitted items may be ambiguous and they are probably best excluded from measures of speededness. Partial attempts, especially those that appear to be truncated, are also ambiguous as noted earlier.

In many ways the exams system for GCSEs, AS and A levels is in a better position here than it was a few years ago. Before marking at item-level became widespread in England, data on not-attempted items was not even available, since a whole-script mark was input by markers. With on-screen item-level marking now being the most frequent marking approach in GCSEs, AS and A levels, item-level data is now available. It should be noted that some online marking systems in use input a ‘0’ for both attempted questions with no rewardable content and non-responses, and this lack of differentiation can limit the analyses that can be carried out.

Some speededness analyses of tests in England have been carried out based on item omission. Wheadon (2011) and Walland (2024) both analysed data from legacy (pre-reform) GCSE tests and found limited evidence of students being affected by test time limits. Wheadon (2011) employed a multi-class mixture Rasch model (Mroch, Bolt and Wollach, 2005) for dichotomous items to investigate the effects of speededness in 5 (pre-reform) foundation tier GCSE papers (2 maths and 3 science). This was done by estimating individual ability on the items at the start of the test and comparing this to ability estimated on items towards the end of the test. Individuals affected by speededness would have a lower ability estimate for the end-of-test items because they had omitted some items they should have been able to answer due to time running out. It was possible from this analysis to determine how many students were affected by the time limit. Results from this study suggested that a substantial proportion of students appeared to have run out of time in one of the maths papers. However, speededness did not appear to have greatly affected students’ performance on the other maths or science papers, except for certain items. One limitation of this approach was that multi-mark items needed to be classified as dichotomous items by classifying low scores as 0 and high scores as 1.

Walland (2024) analysed omitted items at the end of 340 legacy GCSE science and mathematics papers covering the years 2009 to 2016. A continuous string of omitted items at the end of the paper were treated as not reached and the percentage of total paper marks lost for these items was used as a measure of speededness. The students were also stratified by ability to look for differential effects. On average less than 1% of marks were lost on the papers due to not-reached items, with the most significant loss being 4% on one maths paper. Lower ability students lost the most marks, and this indicated that some of the ‘not reached’ items were probably omitted due to their difficulty. Overall it was concluded that most of the exam papers analysed were not speeded, although about 5% of the papers showed 2% or more marks lost on average at the end, indicating some potential speededness around the last item or 2. In summary, these analyses did not suggest that the exams for pre-reform specifications were particularly speeded for the students sitting them.

He and El Masri (2025) carried out an analysis of this kind on 181 current (reformed) specification maths, science and geography GCSE papers from 2017 to 2019. They analysed data which recorded omitted items and concentrated their analysis on those omitted items at the end of the tests which they treated as not reached for time-limit reasons, similarly to Walland (2024). They also applied the partial-credit Rasch model to the data and estimated student ability when not-reached items were treated as missing, so that they did not reduce the ability estimate, but also when the not-reached items were set to a mark of zero so that they lowered the ability estimate. The difference between these 2 estimates was used to classify individuals into different speeded conditions, and to determine how many marks had been lost by estimating what they should have achieved if all items were attempted based on their ability estimated with the not-reached items excluded.

They found variable evidence of speededness across the papers, with most papers appearing to be speeded to a small extent, but with some appearing to be speeded for a larger proportion of students. For example, there were papers where more than 5% of students appeared to lose at least 3% of marks due to not-reached items, particularly on geography and combined science foundation tier papers. Repeating a similar analysis to Walland (2024), He and El Masri (2025) found only slightly higher average marks lost across all the papers. However, looking separately at the 2 tiers, there was little difference between the higher tier papers across the 2 studies, but a substantial, almost two-fold, increase was seen in the marks lost on the foundation tier papers in He and El Masri. It would therefore appear that the foundation-tier papers of the reformed GCSEs may be more speeded than the pre-reform equivalents. These effects were mostly of the order or 0.5% to 2% of marks lost. However, some individual students appeared to have lost substantially more marks, although disentangling time and difficulty effects remains difficult in these more demanding end-of-test items. In contrast to Walland, He and El Masri’s analysis using the Rasch model found more evidence of speededness for moderate ability students, with low-and high-ability students showing much less evidence of the impact of time limits on their marks.

Finally, it is worth noting that some tests in vocational and technical qualifications are taken online, and many of these contain a similar mix of low-tariff questions to the maths and science papers discussed above. They may include both selected-response and constructed-response items. Assuming full item timing data is collected, some of the speededness analysis approaches described earlier that utilise item-timing information might be applied to these tests. However, 2 limitations to this kind of analysis may be present. In many cases the number of test takers may be relatively low, limiting the power of the models and the quality of their fit to the available data. The second limitation is that as discussed at length earlier, the meaning of time per item is less clear for constructed-response questions than it is for selected-response questions, making the interpretation of models testing for speededness more difficult.

Tests with a few high-tariff items

For essay-based subjects, existing single-administration methods of detecting speededness can tell us little about the speededness of these tests. Completely omitted items are much less likely, and because test takers are more likely to tackle the small number of questions out of numerical sequence it is not possible to classify any not-attempted items as not reached for time reasons. With the degree of current test preparation, questions may be answered in a strategic sequence that has been practiced or taught, or test takers may work out a sequence when they first see the questions, starting with those they believe they can accrue marks most easily on, perhaps due to their content knowledge and their own revision. Following this logic, omitted items are as likely to indicate limited knowledge as limited time.

Item difficulty itself is also less meaningful for extended-response questions than short-answer questions, therefore controlling for item difficulty in an analysis would not be effective. This is because of the distinction between differentiation by task, where question difficulty varies and test-taker ability determines which questions are answered correctly, and differentiation by outcome, where test-taker ability determines the quality of response to a question that all should be able to answer, to some extent. Essays fall into the latter category. Although the difficulty of questions may vary, it is the quality of response that is aimed for that largely determines difficulty.

Fundamentally, without direct comparison of different timing conditions, we may never know what impact time limits have had on such tests. Every single response might reflect the effect of time limits or the effect of limited knowledge or ability, or both. A brief, or abruptly-ended, response may well indicate time was not available near the end of the test to finish the response, but it may also indicate abandonment of the answer because the test taker had run out of knowledge or motivation, or intended to return to the item later on but never did. An answer that looks full and complete may not actually be the test taker’s best response. Time pressure (whether real or perceived) may have led them to shorten their planned answer, or spend less time thinking about and planning the response, meaning that a sub-optimal answer is given compared to their maximum potential. For example, some questions separate out marks awarded for knowledge recall or understanding from those awarded for higher-order skills such as analysis and evaluation. When time is limited, a good strategy may be to give a brief, rapid response in the hope that you will at least achieve some of the knowledge marks associated with the item.

Even if item level response time data could be collected, analysing it would also be of limited value for extended-response questions. As already noted, the time taken to write or type the response is often the most significant factor in the total time spent on the question, over and above the thinking and planning time required. Similarly, because of the ambiguities inherent in interpreting responses to extended-response questions, even the more detailed process data that may be collected in CBT such as time spent writing and editing text, or mouse movement and screen scrolling cannot definitively determine if a test taker would have produced a better answer with more time.

For all of the above reasons, even expert scrutiny of individual responses may not yield definitive evidence of speededness. While unanswered questions or partial or truncated responses that appear to be unrelated to question difficulty could suggest that a test is speeded, this will always be ambiguous. Without experimental manipulations, it may always be difficult to say for sure whether a test taker would have scored more highly had more time been available.

Computer-based testing and technology-enhanced items

A final area that should be touched on before thinking about what this all means for testing in England, is the impact that computer-based testing (CBT) may have on the setting of appropriate time limits so that tests are not overly speeded (if at all). Some CBT already occurs in England, and although this is currently not a large part of testing in GCSEs, AS and A levels, its use may grow in future across all qualification types, and so thinking about any impact it may have is useful.

Camara and Harris (2020) describe in detail many of the time considerations that arise from CBT and the kind of items that may be used. While there are various ways CBT may be implemented, and the aim here is not to review this in detail, even the simplest transfer of existing tests with the same type of test items to an on-screen platform (known as ‘paper under glass’) could have some impact on appropriate test timing. There are mode effects in terms of reading time on screen, navigating the question and response space and also inputting the response using different input methods. All of these may alter the time required to answer an item. Some items may be faster to respond to on screen, some faster on paper, and effects will vary between individuals.

However, technology also facilitates the introduction of new item types, such as ones involving embedded video, linked resources or interactive simulations, and these may significantly affect the timing of tests. Such items, known as technology-enhanced items (TEIs), often require more time to complete and may need to be individually trialled to determine reasonable time allowance. This timing may be variable across test takers with different experience of the technology, but also with specific needs, who interact differently with the items. As noted, at a basic level, different input methods (keyboard, touch screen, mouse etc) and response formats (typed words, selections, drag and drop etc) may also change timing needs.

It is clear that careful trialling would be needed in setting appropriate time limits for tests including TEIs, including confirming that existing reasonable adjustments are effective in ensuring a level playing field for those individuals qualifying for them. In addition, any situation where multiple test administration modes ran in parallel (for example CBT and paper-based versions) to assess a common qualification could present significant challenges. Each mode may need different time limits to allow for their different formats, and unless this is agreed and validated to be equivalent across both forms, score comparability across the test forms and technology devices used may be threatened.

Given that the time required to respond to TEIs may vary greatly, any computerised test that selects them from an item bank could have time limits set based on the specific items selected. This could be an overall test duration based on the sum of pre-calculated times for each selected item, or even specific time limits for each item, such that a response to an item has to be given within a time window. If a fixed total test time is required, it would also be straightforward to include pre-defined item response time as an additional constraint for the item-selection algorithm when constructing a test from an item bank.

Tests with different numbers of items for each test taker may require bespoke timing. An example is computer adaptive testing (CAT), where the specific questions presented to test takers are based on their (correct or incorrect) responses to earlier items. The length of CATs vary because the test continues until sufficient measurement precision is achieved (based on an IRT model), not the number of items presented (unless a maximum number of items is specified as a stopping rule).

Bridgeman (2020) points out that care is needed when setting time limits for CATs. Ideally, a generous time would be allowed, because if a strict, fixed test time was used, multiple effects could arise. A slower-working test taker, or one whose pattern of responses leads to a longer adaptive sequence of items, might not reach items that are diagnostic of their true ability. In addition, any tendency towards hasty responses or rapid guessing (for multiple choice items) if time was noted to be running out would have a major impact, as the ability estimate from the CAT algorithm would be reduced by a series of wrong answers on increasingly easy questions towards the end of a test. Finally, more able test takers will be presented with more difficult items, and these may be more complex, with more reading or more complex thought processes required, and so these will typically take longer to respond to than lower-difficulty items. Therefore, either generous timing or a bespoke time based on the items that are presented may be preferable.

CBT also allows process data to be collected if the system allows, which is a more detailed type of data than item response time. It allows different portions of the response process to be identified, such as interacting with a TEI, planning, writing (or responding) and reviewing or editing. Such data may not always be particularly diagnostic of whether tests are actually speeded, because of the complex way test takers may use their time and because the process data itself may be quite hard to analyse and interpret. However, it does potentially offer insight into the way individual items are answered and time is used on the test.

Discussion

In any assessment, a decision needs to be made regarding whether the speed at which an individual works should be part of the assessment construct. In some vocational and technical fields, and some license-to-practice tests, being able to demonstrate fluency such that tasks can be completed within reasonable defined time limits is clearly desirable and relevant to the intended purpose of the qualification. In some cases, the raw speed at which tasks can be completed may be assessed and contribute to outcomes, although this is less common in tests in England than the use of time limits that define a minimum working speed and where marks or outcomes are only based on the correctness or quality of the responses.

For knowledge-based tests, should the speed at which an individual works be considered part of the construct? If knowledge fluency means that a more competent, proficient individual can complete intellectual tasks faster, should that be valued and rewarded? It may be that in some instances, for certain academic subjects, and for some test users, speed in the form of fluency might be valued. Some selection tests such as those for entry into higher education in the US, or job selection tools, may explicitly put test takers under time pressure while carrying out cognitive tasks, and so it may not be an entirely unreasonable idea. This approach has never been openly discussed and debated in England regarding tests of knowledge, skills and understanding, and so it is not clear what desire there may be for incorporating speed into tests. The setting of time limits in assessments has not been researched or discussed widely in the past, and test durations have been assumed to be appropriate and not lead to overly speeded tests.

If there was a consensus that speed of working was important in some cases, there would be 2 key questions to answer. The first is how to reward speed of working in tests of knowledge, skills and understanding, especially those with constructed-response questions, and the second is just how speeded should the assessment be? If knowledge fluency in academic tests were to be valued, careful thought would need to be given to how to design tests that place time limits on the aspects of the response process where speed might be considered important, but do not place time limits on construct-irrelevant parts of the response process, such as writing or typing. Any form of testing on paper is likely to present serious difficulties since applying time restrictions to only specific parts of the response process would not be easy. Clearly, it would be easier to control time use on computer-based tests, and particularly on multiple-choice tests.

A close analysis of the full response process for different types of questions and tasks may be valuable in identifying those aspects for which speed may be considered a valid part of the measurement construct, and those where speed is neutral or definitely construct irrelevant. Such a typology could be useful in stimulating thinking around the types of assessment design that might be appropriate if speed of working were to be assessed.

Second, assuming that test speededness could be measured with reasonable precision, the required degree of speededness would have to be defined. While for a performance assessment there are some work- or task-based expectations of the time tasks should take that can be applied, when demonstrating knowledge and understanding through written questioning it may not be obvious how many test takers should be affected by time limits, and for those that are, how large this impact should be.

This all assumes that speed of working would be integrated into a test through constraining how many items can be answered (or attempted). Those working more quickly would be able to attempt the most items, and so have access to higher marks (ability allowing), while those working more slowly would attempt fewer items which would limit their potential marks. The alternative approach of crediting speed of response directly in scores does not seem as straightforward, but technology could facilitate this through direct measurement of item response time.

If speed were integrated into testing, how individuals that have a disability or other need that affects their speed of working would be accommodated is not clear. Currently, decisions do have to be made about the appropriateness of reasonable adjustments like extra time in assessments that are designed to measure speed of working, such as skills-based performance assessments. In some cases, extra time may be a defensible testing adjustment, while in others the speed of working may be so central to occupational competence that extra time would threaten the validity of the assessment. These are decisions for the awarding organisation developing the qualification to make, in consultation with end-users, and perhaps in some cases wider society, and may depend on the precise occupation for which the test is intended to certify competence. The same would be true in tests of knowledge, skills and understanding where speed of working was part of the test construct. Theoretically, extra time as a “reasonable” adjustment would undermine the purpose of the test. Other reasonable adjustments that are unrelated to time may of course still be appropriate. What this means is that any argument that speed of working was a central part of an assessment needs to be conclusive and well evidenced, and should not undermine the validity of the assessment.

As noted above, if speed were assessed only in certain parts of the response process, this may mitigate unfairness. If particular stages of the response process were separately timed then extra time as a reasonable adjustment might still be appropriate in some stages where it would not threaten the validity of the assessment. Again, this would almost certainly require CBT to allow close timing and control of different aspects of the whole response.

If speed is not made explicit as part of the assessment construct, the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) suggest time limits that do not affect performance are highly desirable. Where time limits impact performance, this may be treated as a threat to test validity, introducing construct-irrelevant variance into test scores. The test outcomes will be affected by both ability (or proficiency) and speed, and as we have seen, these 2 variables are not always closely related. This is because the overall speed at which an individual completes a test of knowledge, skills and understanding across all aspects of the response process is determined by individual dispositional factors as well as their proficiency (and also their writing speed in written assessments).

Measurement of speed of working is not stated as an explicit intention for assessments in any published content[14], conditions, guidance or qualification specification documents for GCSEs, AS and A levels (or indeed many vocational and technical qualifications) in England. Time limits in these exams are set as a convenience, mostly for practical and logistical reasons, although they provide a useful guide as to how long test takers should spend on each question. The implicit intention is that the time limit should be sufficient for most test takers not to be significantly affected by it. In other words they should be able to perform at or near to their maximum in the standard time limits.

Indeed, for true equivalence of test outcomes between those individuals with and without extra time, test takers without extra time should be able to complete the test and mostly demonstrate their best performance in the standard time, and it should only be those with a need that qualifies them for extra time that under-perform under the standard time limits. If standard time limits impact those without extra time, they may end up disadvantaged relative to those receiving extra time.

It remains difficult to know whether the handwritten paper-based exams composed of constructed-response questions in England are in fact speeded. Handwriting speed becomes a major factor in time usage, but a factor which is highly individual and irrelevant to any measurement. It is probably only when less writing is required to respond to the items that an attempt can be made to measure speededness. There is some partial evidence from analysis of not-reached items at the end of science, maths and geography tests that some exams in GCSEs, AS and A levels may be speeded to a degree, although this is not conclusive.

[14] Note though, that while fluency is sometimes mentioned in subject content documents (such as 'read fluently' in GCSE English language) the most direct reference to something like knowledge fluency as discussed in this report comes from the published subject content for GCSE mathematics, which states that "GCSE specifications in mathematics should enable students to develop fluent knowledge, skills and understanding of mathematical methods and concepts" (Department for Education, 2013b, page 3). While this suggests that speed (and ease) of access to this content is valued, there is no explicit stated intent for the actual assessments to measure speed.

Experimental manipulations would probably be needed to more definitively measure speededness, through administering tests under different timing conditions, or allowing the test to continue after standard time and clearly indicating marks that were awarded during or after standard time. However, even these methods are likely to be affected by lower test-taker motivation since they would need to be carried out under low-stakes conditions. It may be possible to survey test takers at the end of such an experimental test and analyse only data from those individuals reporting higher motivation (recognising that such self-reports may not always be reliable). Further thought would be required regarding other potential experimental approaches.

Close scrutiny of responses and professional judgement that they show indications of running out of time cannot be conclusive because of the complex way test takers sequence answering questions and returning to them later, as well as limitations in their knowledge. However, such scrutiny may still be useful in trying to understand whether a perception of time pressure may arise because test takers are trying to write more than they really need to, for example because their responses are unfocussed, with a lot of irrelevant material in their answers. This may be a contributory factor in perceptions of time pressure, especially for the more essay-based subjects.

Even when exams are presented via CBT, with the detailed timing information this can provide, the difficulties in measuring speededness in items requiring constructed responses largely remain since it is the item types themselves that do not lend themselves to straightforward measures of speededness. However, careful analysis of the detailed item timing and process data that CBT allows may provide some insight as to whether time ran out or impacted responses.

It is also worth remembering that with no pre-testing of exam papers taking place in England for most qualifications, the time pressure on different versions of a test may vary. One year may be more speeded than others. In addition, the cohort sitting each version of the test will differ. It may therefore be difficult to reach firm conclusions about whether assessments for a particular qualification are speeded without testing many different versions, even if one test is found to be.

Qualifications are regularly reviewed and reformed, and so discussion and debate around the role of speed of working in different assessments is valuable, allowing a clearer understanding of what is desirable in future tests. If speed were decided to be construct-relevant in some tests of knowledge, skills and understanding, then this would need to be made explicit, and potentially be supported by a validity argument, with thought given to what this would mean for students who currently receive extra time. How to integrate speed of working into the measurement model and indeed how much speededness to incorporate would need to be carefully thought through. In any future transition to a more technology-enabled testing context, time limits would need to be researched for different types of items. The potential to apply time restrictions to just some parts of the task response process could also be investigated.

If speed were decided to be construct-irrelevant for these tests of knowledge, skills and understanding, then future assessment development would need to start with the explicit goal of making sure testing time was sufficient for most test takers to complete tests, rather than this being an implicit goal that is assumed to be met and not carefully investigated. Further work would be valuable in evaluating the speededness of current exams in GCSEs, AS and A levels, given some partial evidence of speededness exists. This may include exploring the use of experimental manipulations of testing conditions, test data analysis for some types of test, and response scrutiny for more extended-response questions.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for Educational and Psychological Testing.* American Educational Research Association. Retrieved from https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

AQA. (n.d.). Making an exam - a guide to creating a question paper. Retrieved from https://www.youtube.com/watch?v=nQcSXv6PcXs

Association of Educational Assessment - Europe (2022). *European Framework of Standards for Educational Assessment 1.0.* Edizioni Nuova Cultura. Retrieved from https://aea-europe.net/wp-content/uploads/2022/11/SW_Framework_of_European_Standards.pdf

Becker, B., Weirich, S., Goldhammer, F. and Debeer, D. (2023). Controlling the Speededness of Assembled Test Forms: A Generalization to the Three-Parameter Lognormal Response Time Model. *Journal of Educational Measurement*, *60*, 551-574. https://doi.org/10.1111/jedm.12364

Benton, T. (2017). How much do I need to write to get top marks? *Research Matters: A Cambridge Assessment publication*, *24*, 37-40. https://doi.org/10.17863/CAM.100359

Binder, C. (1996). Behavioral fluency: Evolution of a new paradigm. *The Behavior Analyst*, *19*(2), 163-197. https://doi.org/10.1007/BF03393163

Biola, H. R. (1982). Time limits and topic assignments for essay tests. *Research in the Teaching of English*, *16(1)*, 97–98. https://www.jstor.org/stable/40170881

Bolsinova, M., Tijmstra, J., & Molenaar, D. (2017). Response moderation models for conditional dependence between response time and response accuracy. *British Journal of Mathematical & Statistical Psychology*, *70*, 257–279. https://doi.org/10.1111/bmsp.12076

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*, 331–348. https://doi.org/10.1111/j.1745-3984.2002.tb01146.x.

Bridgeman, B. (2020). Relationship between testing time and testing outcome. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating Timing Considerations to Improve Testing Practices* (pages 59–72). Routledge.

Bridgeman, B., Trapani, C., & Curley, E. (2004). Impact of fewer questions per section on SAT I scores. *Journal of Educational Measurement*, *41*(4), 291–310. https://doi.org/10.1111/j.1745-3984.2004.tb01167.x

Cahan, S., Nirel, R., & Alkoby, M. (2016). The extra-examination time granting policy: A reconceptualization. *Journal of Psychoeducational Assessment*, *34*(5), 461–472. https://doi.org/10.1177/0734282915616537

Camara, W. J. & Harris, D. J. (2020). Impact of technology, digital devices, and test timing on score comparability. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating Timing Considerations to Improve Testing Practices* (pages 104–121). Routledge.

Caudery, T. (1990). The validity of timed essay tests in the assessment of writing skills. *ELI Journal*, *44*, 122–131. https://doi.org/10.1093/elt/44.2.122

Chase, W. G. and Simon, H. A. (1973). The mind's eye in chess. https://doi.org/10.1016/B978-0-12-170150-5.50011-1. In W. G. Chase (Ed.) *Visual Information Processing*. Academic Press.

Chi, M. T. H., Glaser, R., and Rees, E. (1982). Expertise in problem solving. In R. J. Sternberg (Ed.) *Advances in the Psychology of Human Intelligence, Vol. 1*. Erlbaum.

Cintron, D. W. (2021). Methods for measuring speededness: Chronology, classification, and ensuing research and development. ETS Research Report Series RR-21-22, Educational Testing Service. https://doi.org/10.1002/ets2.12337

Cizek, G. (2005). More Unintended Consequences of High-Stakes Testing. *Educational Measurement: Issues and Practice*, *20*, 19-27. https://doi.org/10.1111/j.1745-3992.2001.tb00072.x

Clauser, B. E., Margolis, M. J., & Clauser, J. C. (2017). Validity issues for technology-enhanced innovative assessments. In H. Jiao & R. W. Lissitz (Eds.), *Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective* (pages 139–161). Information Age Publishing.

Cronbach, L. J., & Warrington, W. G. (1951). Time limit tests: Estimating their reliability and degree of speeding. *Psychometrika*, *14*, 167–188. https://doi.org/10.1007/BF02289113

Cuff, B.M.P., Keys, E., Churchward, D., Kennedy, L., & Holmes, S. (2025). *Extra time in assessments: A review of the research literature on the effect of extra time on assessment outcomes for different types of student*. (Ofqual research report 25/7267/1) Ofqual. Retrieved from https://www.gov.uk/government/publications/extra-time-in-assessments

Davidson, W. M., & Carroll, J. B. (1945). Speed and level components in time-limit scores: A factor analysis. *Educational and Psychological Measurement*, *5*(4), 411–427. https://doi.org/10.1177/001316444500500408

De Boeck, P. & Rijmen, F. (2020). Response times in cognitive tests: Interpretation and importance. In M. J. Margolis & R. A. Feinberg (Eds.), Integrating Timing Considerations to Improve Testing Practices (pages 142–149). Routledge.

de Groot, A. (1966). Perception and memory versus thought: Some old ideas and recent findings. In B. Kleinmuntz (Ed.) *Problem Solving*. Wiley.

Department for Education (2013a). *English language: GCSE Subject Content and Assessment Objectives*. Department for Education. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/254497/GCSE_English_language.pdf

Department for Education (2013b). *Mathematics: GCSE Subject Content and Assessment Objectives*. Department for Education. Retrieved from https://assets.publishing.service.gov.uk/media/5a7cb5b040f0b6629523b52c/GCSE_mathematics_subject_content_and_assessment_objectives.pdf

Department for Education (2025). *Curriculum and assessment review final report: Building a world-class curriculum for all.* Department for Education. Retrieved from https://www.gov.uk/government/publications/curriculum-and-assessment-review-final-report

Evans, F. R., & Reilly, R. R. (1972). A study of speededness as a source of test bias. *Journal of Educational Measurement*, *9*, 123–131. https://doi.org/10.1111/j.1745-3984.1972.tb00767.x

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223-241. https://doi.org/10.1177/1745691612460685

Feinberg, R. A., & Jurich, D. (2018). *Using rapid responses to evaluate test speededness*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY, Apr 2018.

Feinberg, R., Jurich, D., & Wise, S.L. (2021). Reconceptualizing rapid responses as a speededness indicator in high-stakes assessments, *Applied Measurement in Education*, *34*(4), 312-326, https://doi.org/10.1080/08957347.2021.1987904

Gernsbacher, M.A., Soicher, R.N., & Becker-Blease, K.A. (2020). Four empirically based reasons not to administer time-limited tests. *Translational Issues in Psychological Science*, *6*(2), 175–190. https://doi.org/10.1037/tps0000232

Goldhammer, F., Naumann, J., Stelter, A., Toth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*, 608–626. https://doi.org/10.1037/a0034716

Gulliksen, H. (1950). *Theory of mental tests*. John Wiley and Sons. https://doi.org/10.1037/13240-000

Hale, G. (1992). *Effects of the amount of time allowed on the test of written English* (Research Report RR-92-27). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1992.tb01458.x

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer Academic.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.

Harik, P., Clauser, B. E., Grabovsky, I., Baldwin, P., Margolis, M. J., Bucak, D., Jodoin, M., Walsh, W. and Haist, S. (2018). A comparison of experimental and observational approaches to assessing the effects of time constraints in a medical licensing examination. *Journal of Educational Measurement*, *55*(2), 308–327. https://doi.org/10.1111/jedm.12177

Harik, P., Feinberg, R. A., & Clauser, B.E. (2020). How examinees use time: Examples from a medical licensing exam. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating Timing Considerations to Improve Testing Practices* (pages 32–46). Routledge.

He, Q. and El Masri, Y. (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

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*, article 150. https://doi.org/10.3389/fnins.2014.00150

Henmon, V. A. C.(1911).The relation of the time of a judgment to its accuracy. *Psychological Review*, *18*, 186–201. https://doi.org/10.1037/h0074579

Himmelweit, H. T. (1946). Speed and accuracy of work as related to temperament. *British Journal of Psychology*, *36*, 132–144. https://doi.org/10.1111/j.2044-8295.1946.tb01115.x

Jerrim, J. (2023). Test anxiety: Is it associated with performance in high-stakes examinations?, *Oxford Review of Education*, *49*(3), 321-341, https://doi.org/10.1080/03054985.2022.2079616

Joint Council for Qualifications (2019). *Report of the Independent Commission on Exam Malpractice*. Joint Council for Qualifications. Retrieved from https://www.jcq.org.uk/wp-content/uploads/2019/09/Independent_Report.pdf

Jurich, D.P. (2020). A history of test speededness: tracing the evolution of theory and practice. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating Timing Considerations to Improve Testing Practices* (pages 1–18). Routledge.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kane, M. (2020). The impact of time limits and timing information on validity. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating Timing Considerations to Improve Testing Practices* (pages 19–31). Routledge.

Kennedy, M. (1930). Speed as a personality trait. *The Journal of Social Psychology*, *1*, 286–299. https://doi.org/10.1080/00224545.1930.9918819

Klein, S. P. (1981). *The effect of time limits, item sequence, and question format on applicant performance on the California Bar Examination.* A Report submitted to the Committee of Bar Examiners of the state of California and the National Council of Bar Examiners.

Koretz, D.M., Linn, R.L., Dunbar, S.B., & Shepard, L.A. (1991) *The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests*. Paper presented in R.L. Linn (Chair), Effects of High-Stakes Educational Testing on Instruction and Achievement, symposium presented at the annual meeting of the American Educational Research Association, Chicago, April 5, 1991.

Leckie, G. and Goldstein, H. (2017). The evolution of school league tables in England 1992–2016: ‘Contextual value-added’, ‘expected progress’ and ‘progress 8’. *British Educational Research Journal*, *43*: 193-212. https://doi.org/10.1002/berj.3264

Li, D., Yi, Q., & Harris, D. (2016). *Evidence for paper and online ACT comparability* (ACT Working Paper 2016-02). ACT. Retrieved from https://www.act.org/content/dam/act/unsecured/documents/Working-Paper-2016-02-Evidence-for-Paper-and-Online-ACT-Comparability.pdf

Lovett, B. J. (2010). Extended time testing accommodations for students with disabilities: Answers to five fundamental questions. *Review of Educational Research*, *80*(4), 611–638. https://doi.org/10.3102/0034654310364063.

Lovett, B. J. (2020). Extended time testing accommodations for students with disabilities: Impact on score meaning and construct representation. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating Timing Considerations to Improve Testing Practices* (pages 47–58). Routledge.

Lu Y., & Sireci S.G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice*, *26*, 29–37. https://doi.org/10.1111/j.1745-3992.2007.00106.x

Luce, D. (1986). *Response Times: Their Role in Inferring Elementary Mental Organization*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195070019.001.0001

Margolis, M. J., von Davier, M. & Clauser, B. E. (2020). Timing considerations for performance assessments. In M. J. Margolis & R. A. Feinberg (Eds.), Integrating Timing Considerations to Improve Testing Practices (pages 90–103). Routledge.

Margolis, M. J. & Feinberg, R. A. (Eds.). *Integrating Timing Considerations to Improve Testing Practices*. Routledge. Available at https://www.taylorfrancis.com/books/oa-edit/10.4324/9781351064781/integrating-timing-considerations-improve-testing-practices-melissa-margolis-richard-feinberg

Morin, C., Black, Howard, E, Holmes, S.D. (2018). *A study of hard to mark responses: Why is there low mark agreement on some responses?*. (Ofqual research report 18/6449/5). Ofqual. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/759216/HardtoMark_-_FINAL64495.pdf

Mollenkopf, W. G. (1950). An experimental study of the effects on item-analysis data of changing item placement and test time limit. *Psychometrika*, *15*, 291–315. https://doi.org/10.1007/bf02289044

National Board of Osteopathic Medical Examiners, Inc. (2013). *Knowledge fluency and time limitations*. Unpublished manuscript.

Mroch, A. A., Bolt, D. M. & Wollack, J. A. (2005). *A new Multi-Class Mixture Rasch Model for test speededness*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada, Apr. Retrieved from https://testing.wisc.edu/research%20papers/NCME%202005%20paper%20(Mroch,%20Bolt,%20&%20Wollack).pdf

Nisbett, I. & Shaw, S. D. (2020). *Is Assessment Fair?* Sage.

Ofqual (2015). *GCSE subject-level conditions and requirements for English language and certificate requirements (Version 4 – July 2015)*. Ofqual. Retrieved from https://www.gov.uk/government/publications/gcse-9-to-1-subject-level-conditions-and-requirements-for-english-language

Oppler, S. H., Davies, S. A., Lyons, B. D., Nathanson, L. B., & Chen,W. (2004). *The effect of speededness on MCAT scores: An initial examination*. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL, Apr 2004.

Putwain, D. and Daly, A. L. (2014). Test anxiety prevalence and gender differences in a sample of English secondary school students, *Educational Studies*, *40*(5), 554-570, https://doi.org/10.1080/03055698.2014.953914

Rindler, S. E. (1979). Pitfalls in assessing test speededness. *Journal of Educational Measurement*, *16*, 261–270. https://doi.org/10.1111/j.1745-3984.1979.tb00107.x

Robin, F., & Zhao, J. C. (2014). Timing of the analytic writing measure of the GRE revised general test. In C. Wendler & B. Bridgeman (Eds.), *The research foundation for the GRE revised general test: A compendium of studies* (pages 1.8.1–1.8.8). Educational Testing Service.

Schnipke, D. L. (1995). Assessing speededness in computer-based tests using item response times. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA, Apr 1995.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*, 213–232. https://doi.org/10.1111/j.1745-3984.1997.tb00516.x

Shao, C., Li, J., & Cheng, Y. (2016). Detection of test speededness using changepoint analysis. *Psychometrika*, *81*(4), 1118–1141. https://doi.org/10.1007/s11336-015-9476-7

Sireci, S.G. (2020). Standardization and UNDERSTANDardization in Educational Assessment, *Educational Measurement: Issues and Practice*, *39*(3), 100–105. https://doi.org/10.1111/emip.12377

Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, *75*(4), 457–490. https://doi.org/10.3102/00346543075004457

Spalding, V. (2010). *Structuring and formatting examination papers: Examiners' views of good practice*. AQA Centre for Education Research and Policy. Retrieved from https://filestore.aqa.org.uk/content/research/CERP-RP-VS-05022010_0.pdf?download=1 .

Spearman, C. (1927). *The Abilities of Man*. Macmillan.

Stafford, R. E. (1971). The speed quotient: A new descriptive statistic for tests. *Journal of Educational Measurement*, *8*, 275–278. https://doi.org/10.1111/j.1745-3984.1971.tb00937.x

Swanson, D. B., Case, S. M., Ripkey, D. R., Clauser, B. E., & Holtman, M. C. (2001). Relationships among item characteristics, examinee characteristics, and response times on the USMLE Step 1. *Academic Medicine*, *76*(10), S114–S116. https://doi.org/10.1097/00001888-200110001-00038

Swineford, F. (1956). *Technical manual for users of test analyses (SR-56-42)*. Educational Testing Service.

Swineford, F. (1974). *The test analysis manual* (SR-74-06). Educational Testing Service.

Tate, M. W. (1948). Individual differences in speed of response in mental test materials of varying degrees of difficulty. *Educational and Psychological Measurement*, *8*(3-1), 353–374. https://doi.org/10.1177/001316444800800307

Thompson, S., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). University of Minnesota, National

Center on Educational Outcomes. Retrieved from https://nceo.umn.edu/docs/OnlinePubs/Synth44.pdf

Thorndike, E. L., Bregman, E. O., Cobb, M. V., & Woodyard, E. (1926). *The measurement of intelligence*. Teachers College Bureau of Publications. https://doi.org/10.1037/11240-000

Thurstone, L.L. Ability, motivation, and speed. *Psychometrika*, *2*, 249–254 (1937). https://doi.org/10.1007/BF02287896

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308. https://doi.org/10.1007/s11336-006-1478-z

van der Linden, W. J. (2011). Test design and speededness. *Journal of Educational Measurement*, *48*(1), 44–60. https://doi.org/10.1111/j.1745-3984.2010.00130.x

van der Linden, W. J. (2017). Test speededness and time limits. In W. J. van der Linden (Ed.) *Handbook of item response theory, volume three: Applications* (pages 249-266). Chapman and Hall/CRC. https://doi.org/10.1201/9781315117430

Walland, E. (2024). Exploring speededness in pre-reform GCSEs (2009 to 2016). *Research Matters: A Cambridge University Press & Assessment Publication*, *37*, 57–73. https://doi.org/10.17863/CAM.106035

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*(1), 67–85. https://doi.org/10.1016/0001-6918(77)90012-9

Wheadon, C. (2011). *An Item Response Theory approach to the maintenance of standards in public examinations in England* (Doctoral Thesis, Durham University). Retrieved from https://etheses.dur.ac.uk/615/1/Chris_Wheadon_PhD.pdf?DDD29

Wild, C. L., Durso, R., & Rubin, D. B. (1982). Effect of increased test-taking time on test scores by ethnic group, years out of school, and sex. *Journal of Educational Measurement*, *19*, 19–28. https://doi.org/10.1111/j.1745-3984.1982.tb00111.x

Wise, S. L. (2015). Response time as an indicator of test taker speed: Assumptions meet reality. *Measurement: Interdisciplinary Research & Perspective*, *13*, 186–188. https://doi.org/10.1080/15366367.2015.1105062

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*, 19–38. https://doi.org/10.1111/j.1745-3984.2006.00002.x

Wise, S., & Kuhfeld, M. (2021). A method for identifying partial test-taking engagement. *Applied Measurement in Education*, *34*(2), 150-161. https://doi.org/10.1080/08957347.2021.1890745

Yamamoto, K. (1995). *Estimating the effects of test length and test time on parameter estimation using the HYBRID model* (TOEFL Tech. Rep. No. TR-10). Educational Testing Service. Retrieved from https://doi.org/10.1002/j.2333-8504.1995.tb01637.x

Yerkes R.M. & Dodson J.D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, *18* (5): 459–482. https://doi.org/10.1002/cne.920180503

Zuriff, G. E. (2010) Extra Examination Time for Students With Learning Disabilities: An Examination of the Maximum Potential Thesis, *Applied Measurement in Education*, *13*(1), 99-117. https://doi.org/10.1207/s15324818ame1301_5

Published by:

ofqual

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

0300 303 3344

public.enquiries@ofqual.gov.uk

www.gov.uk/ofqual

November 2025

Ofqual/25/7267/2