

Rapid systematic evidence review on summative assessment in upper secondary education: A review across ten jurisdictions



Alison O'Mara-Eves, Antonia Simon, Sum Yue (Jessica) Ko, and Kusha Anand

Table of Contents

Table of Contents	2
Abstract.....	5
The authors of this report are:	6
Acknowledgements.....	6
About the EPPI Centre	6
About the funder	6
Conflicts of interest.....	6
Contributions.....	6
Citation	6
One page summary of key findings.....	7
Key findings about purposes of the assessment.....	7
Key findings about the study outcomes	7
Key findings across research and policy sources	8
Executive summary.....	9
Introduction.....	9
Methods	9
Results of thematic summaries and thematic synthesis	10
Results of comparative synthesis	11
Main report	12
Introduction.....	12
Background	12
The commissioning context and broader project.....	12
Methods for the research evidence review	14
Research question	14
Design	15
Study identification	15
Data Extraction	17
Critical Appraisal.....	18
Synthesis	18
Methods for the cross-jurisdictional comparison table	22
Results of the thematic summary	22
Key conclusions from the thematic summary	36
Results of the thematic synthesis	36

Themes about purposes of the assessment	36
Progression to the next stage of education	36
Student learning	37
Quality assurance.....	37
International comparisons.....	38
Fairness and equality.....	38
Summary of findings about purposes of the assessment	39
Themes about the study outcomes	40
Attainment	40
Student motivation	40
Student stress and anxiety.....	41
Summary of findings about the study outcomes	41
Debates in the literature on summative assessment.....	42
Potential harms to students	42
Implications for teaching practice	42
Implications for future assessment.....	42
Reliability and validity of assessments.....	43
Predictability of content within assessments.....	43
High stakes assessment	43
Relevance of disadvantage	43
Student voice in policy reform	43
Breadth versus depth.....	44
Summary of the debates identified in the literature.....	44
Conceptual framework of summative assessment	44
Results of the comparative synthesis.....	47
Results of the comparison tables.....	49
Discussion	67
Summary of findings across the research	67
Description of the evidence on summative assessment in general lower and upper secondary education (ages 14-18/19)	67
Summary of findings about purposes of the assessment	67
Summary of findings about the study outcomes	68
Summary of the debates identified in the literature.....	68
Summary of findings from the comparative synthesis.....	69
Transferability to the English context	69
Strengths and limitations of the review	70

Deviations from proposal	71
Conclusion.....	71
References included in the review (n = 21).....	73
Appendix A. NFER Scoping study.....	75
Appendix B. Methods for the review of policy sources.....	76
Appendix C. Data extraction tool	78
Appendix D. Study findings about progression to the next stage of education	82
Appendix E. Study findings about student learning	85
Appendix F. Study findings about equality and fairness within the education system.....	90
Appendix G. Study findings about student attainment	93
Appendix H. Study findings about student motivation	102
Appendix I. Study findings about student stress and anxiety.....	107

Abstract

We conducted a rapid systematic evidence synthesis of 21 studies on summative assessment in general lower and upper secondary education (ages 14-18/19) in 10 jurisdictions (England, Estonia, Finland, Hong Kong, Japan, the Netherlands, New Zealand, Ontario - Canada, Poland, and Republic of Ireland). A thematic synthesis found that assessment systems generally achieve the purposes of allowing progression to the next stage of education and assessing student learning. It was unclear whether assessment policies support quality assurance or the accountability of schools or the education system. Major concerns were raised about ensuring fairness and equality, which was not always met. Regarding student outcomes, assessment systems have a complex relationship with student attainment, and no definitive conclusions were drawn about their impact on student motivation. Increased stress and anxiety outcomes were commonly (although not universally) reported, particularly in high-stakes contexts. Compared with themes identified across the policy sources, the research sources tend to focus on individual-level outcomes for students or teachers, whereas policy sources tend to focus on national- or international-level concerns. A key area of overlap was the desire for fairness and inclusivity for students, which does not always appear to be met; the involvement of student voice in policy reform was highlighted as an important strategy for addressing this.

The authors of this report are:

Alison O'Mara-Eves^{1*}, Antonia Simon^{1*}, Sum Yue (Jessica) Ko¹, and Kusha Anand¹

¹EPPI Centre, UCL Social Research Institute; *Joint first authors

Contact: a.omara-eves@ucl.ac.uk or a.simon@ucl.ac.uk

Acknowledgements

We thank Katie Hollingworth for her work on the individual policy briefs we drew on when developing the policy comparison table. We thank Patrick Wall from the Centre for Education Systems (CES) for comments on draft versions, and Loic Menzies from CES for commenting on the policy comparison table. Front cover image cropped from a photo by Mary Hinkley, 2018, © UCL Digital Media.

About the EPPI Centre

The Evidence for Policy and Practice Information Centre is based at the Social Science Research Unit within the UCL Social Research Institute, a department of IOE, UCL's Faculty of Education and Society. The EPPI Centre's mission is to promote and support collaborative, rigorous evidence for a fairer world by focusing on evidence synthesis methods, participation, research impact, and capacity building. For more information, visit <https://eppi.ioe.ac.uk/>.

About the funder

The organisation edpol.net (now the Centre for Education Systems, CES) commissioned this pilot review to assess the feasibility of both working with recognised academic specialists and to create a replicable review process for the wider CES programme. CES is committed to reviewing all areas of system policy making, providing descriptions and insight from the four UK Home Nations, plus ten other jurisdictions around the world. As their insight and comparative analysis builds, they will provide a critical resource for policy makers and influencers, including alternative approaches for individual policy initiatives, an increased understanding of how policy instruments interact and a system architecture that identifies investment priorities.

Conflicts of interest

There were no conflicts of interest in conducting this research or the writing of this report.

Contributions

Responsibility for the views expressed remains solely with the authors. This report is an early-stage proof-of-concept study for Charitable Foundation for Education Development and edpol.net. Any comments or insights are the authors' and do not necessarily express the views of the funding and management organisations.

Citation

The report should be cited as: O'Mara-Eves A, Simon A, Ko S Y (J), and Anand K. (2024). *Rapid systematic evidence review on summative assessment in upper secondary education: A review across ten jurisdictions*. London: EPPI Centre, Social Science Research Unit, UCL Institute of Education, University College London.

© Copyright 2024 Authors of the systematic reviews on the EPPI Centre website (<http://eppi.ioe.ac.uk/>) hold the copyright for the text of their reviews. The EPPI Centre owns the copyright for all material on the website it has developed, including the contents of the databases, manuals, and keywording and data-extraction systems. The centre and authors give permission for users of the site to display and print the contents of the site for their own non-commercial use, provided that the materials are not modified, copyright and other proprietary notices contained in the materials are retained, and the source of the material is cited clearly following the citation details provided. Otherwise, users are not permitted to duplicate, reproduce, re-publish, distribute, or store material from this website without express written permission.

One page summary of key findings

- The review included 21 primary research studies from seven jurisdictions. These were predominantly studies from England (n = 12) and the Republic of Ireland (n = 7). This was followed by Finland, Hong Kong, and New Zealand (n = 2 each); and Ontario and The Netherlands (n = 1 each). No eligible studies were identified for Estonia, Poland, or Japan.
- Summative assessment is a highly contentious topic, with debates on diverse aspects, including potential harms to students; issues around equity, fairness, and disadvantage; the form and function of the assessments; and the coverage of knowledge and skills.

Key findings about purposes of the assessment

- Assessment systems achieve the purpose of allowing progression to the next stage of education and were generally found to capture individual student learning well.
- There was insufficient evidence to determine whether the assessment policies enable quality assurance or auditing of the quality of schools or the education system.
- The purpose of ensuring fairness and equality was not always met, with concerns raised about gender stereotyping, teacher bias, subject bias, difference between socioeconomic groups, and inequalities in the grading structures.
- Concerns were also raised about whether the assessment systems:
 - equipped students for the future, particularly practical skills.
 - overemphasised maximising grades at the expense of learning.
 - risked limiting opportunities if students failed to pass threshold grades.
 - become overly complex when both formative and summative elements are used.
 - were adequately explained to the students.
 - unfairly prioritised certain academic subjects (e.g., maths) over others (e.g., art).
- Facilitators to improved learning and successful assessment outcomes included:
 - providing helpful, relevant feedback
 - providing content specific to the subject being taught
 - including student voice when designing or revising secondary level assessments.

Key findings about the study outcomes

- Assessments have a complex relationship with students' attainment.
- Student motivation varied depending on the system features, the purpose of the test, and the individual student. Teachers are likely to be important facilitators of student motivation.
- Stress and anxiety were commonly reported, especially for high-stakes assessments, although a few studies did not find this negative association.

Key findings across research and policy sources

- The research sources tend to focus on individual level outcomes for students or teachers, such as student anxiety (bottom-up), whereas policy sources tend to focus on national or international-level concerns (top-down), such as labour market needs.
- A key area of overlap between research and policy sources was around the desire for fairness and inclusivity for students, which appears to not always be met in practice; involvement of student voice in policy reform was highlighted as an important strategy for addressing this.

Executive summary

Introduction

Summative assessments are tests or exams that measure students' learning at the end of a course or a school year. They are widely used in education systems around the world, but they also raise many questions and debates about their purposes, effects, and implications.

We conducted a rapid systematic evidence synthesis of studies on summative assessment in general lower and upper secondary education (ages 14-18/19) in 10 jurisdictions: England, Estonia, Finland, Hong Kong, Japan, the Netherlands, New Zealand, Ontario - Canada, Poland, and Republic of Ireland. The review aimed to inform discussions about policy development, improvement, and reform in England, where summative assessment policies have undergone several changes in recent years.

Methods

The review followed a standard systematic review methodology, with some adaptations to suit the rapid nature of the project. The review question was:

What does academic research tell us about the extent to which the assessment systems meet the purposes stated in the primary studies? What does rigorous evidence tell us about the impact of assessment systems and policy on students' motivation to learn and attainment?

The review included primary research studies that met the following criteria:

- Published between 2012 and 2022.
- Reported on assessment systems and policy made at the jurisdiction level (i.e., national or state), relevant to the main summative assessment pathways in general secondary education (academic and/or technical).
- Focused on general lower and upper secondary education (ages 14-18/19)
- Conducted in one of the 10 selected jurisdictions.
- Reported empirical data on student outcomes (attainment, motivation, stress and anxiety) or purposes of summative assessment policies.
- Written in English
- Reported large-scale primary research (qualitative and quantitative), secondary analysis, or systematic literature review/ cross-jurisdiction map.

The review team extracted data from each study using a tool that covered information on assessment characteristics, study characteristics, methods, findings, debates in the literature, and critical appraisal.

The synthesis occurred in three main stages: thematic summaries, thematic synthesis, and a comparative synthesis that ties in the findings from the policy documents. Thematic summaries produce a descriptive account of the evidence (Thomas, O'Mara-Eves, Harden,

Newman, 2017)¹. The thematic synthesis explores themes identified in the studies and develops new theories or frameworks by exploring points of convergence and difference amongst the studies (Thomas et al., 2017). The comparative synthesis brings together our understanding of summative assessment that we have gained through exploring the research evidence (the focus of this report) and what we have learned through examining the policy sources (the focus of individual jurisdiction reports). Finally, a comparison of policies across jurisdictions are presented in tabular form in the comparative table section of the report.

Results of thematic summaries and thematic synthesis

The review included 21 primary research studies from seven jurisdictions. These were predominantly studies from England (n = 12) and the Republic of Ireland (n = 7). This was followed by two each from Finland, Hong Kong, and New Zealand, and one each from Ontario and the Netherlands. No eligible studies were found for Estonia, Poland, and Japan.

Through the thematic summaries, we identified that the most reported *purpose* of the assessment system was to assess whether learning objectives have been achieved (n = 9), followed by enabling progression through levels of education (n = 8). In contrast, the most reported *use* of the assessment system was to enable progression through levels of education (n = 14), with the majority of those focused on progression to higher education (n = 12). The next most common use was to allow employers to differentiate between candidates (n = 11). Most of the studies explicitly stated that the assessment used examinations to assess the required standard (n = 15), sometimes in combination with other assessment types such as coursework.

The results of the thematic synthesis suggest that summative assessment systems generally achieve the purposes of allowing progression to the next stage of education and assessing student learning. However, there were major concerns raised about fairness and equality for students, which was not always met by the assessment policies. For example, some studies reported gender stereotyping, teacher bias, subject bias, socioeconomic differences, and grading inequalities in relation to summative assessment. Some studies also questioned whether the assessment systems equipped students for the future.

Summative assessment systems appear to have a complex relationship with student outcomes including attainment, motivation, stress, and anxiety. These outcomes may vary depending on the features of the assessment system, the purpose of the test, and the individual student. Some studies reported positive effects of summative assessment on student outcomes, such as increased motivation, improved learning, and enhanced performance. Other studies reported negative effects, such as increased stress and anxiety, reduced motivation, and distorted learning.

The review identified several debates and challenges in the literature on various aspects of summative assessment, including potential harms to students, implications for teaching practice, reliability and validity of assessments, predictability of content, high stakes

¹ Thomas, J, O'Mara-Eves, A, Harden, A, & Newman, M. (2017). Synthesis methods for combining and configuring textual or mixed methods data. In D Gough, S Oliver, and J Thomas (eds.), *An introduction to systematic reviews* (2nd ed.), London: SAGE.

assessment, relevance of disadvantage, student voice in policy reform, and breadth versus depth of knowledge and skills.

Results of comparative synthesis

The comparative synthesis identified that there is a gap between the research and policy sources in terms of their focus and perspective. The research sources tend to focus on individual level outcomes for students or teachers, such as student anxiety, motivation, or attainment, whereas policy sources tend to focus on national or international-level concerns, such as labour market needs, quality assurance, or international comparisons.

A key area of overlap between research and policy sources is the desire for fairness and inclusivity for students, which appears to not always be met in practice. Some jurisdictions have implemented reforms to address this issue, such as introducing more flexibility and choice for students, reducing high-stakes testing, or involving student voice in policy development and implementation.

The jurisdictions vary in their balance of formative and summative assessment, with some jurisdictions relying more on external examinations and others using more continuous assessment or a combination of both. The jurisdictions also differ in their timing and extent of assessment policy reform, with some jurisdictions undergoing several major changes in recent years and others maintaining more stability or continuity.

The comparison tables provide a descriptive analysis 'at a glance' of the key policy differences and similarities between the jurisdictions. The key results mirror a lot of the findings mentioned above. A key point is that every jurisdiction has its own ways of doing things, although there are some parallels. The parallels are particularly evident between the EU countries, in terms of aspirations to deliver a system that values formative feedback and recognises student individuality. While it was difficult to determine whether a given jurisdiction has high, medium, or low stakes from the perspective of the student, most systems offered a mixture of different assessment forms, typically culminating in a final examination. The emphasis placed on the final exam may differ; for example, England places heavy emphasis on the final exams, whereas Finland, the Republic of Ireland, Ontario, and New Zealand seem to give more relative weight to other assessment types.

Main report

Introduction

Background

Summative assessments in general lower and upper secondary education (ages 14-18/19) are prevalent in education systems across the world. According to the jurisdiction report for England,

Assessment practises and policies [in England] have evolved over time, with adjustments made to the types of assessments used and how they are delivered. The current summative assessment system in the United Kingdom comprises modifications to the General Certificate of Secondary Education (GCSE), Advanced Subsidiary (AS) level, and Advanced (A) level qualifications. The assessment system reforms were first announced in 2013, and they were gradually implemented beginning in 2015. These reforms attempt to create a more academically challenging curriculum focused on preparing pupils for further education, higher education, and jobs. Furthermore, subject-specific adjustments, such as enhanced breadth and depth of information and new themes and concepts, were included in the reforms. These changes were implemented as part of a larger attempt to improve the rigour and quality of the UK education system, as well as to ensure that qualifications are better connected with the needs of employers and further education institutions.

...Problems with the current system have been noted, such as a lack of congruence between the curriculum and exams. Further, contextual variables such as changing educational demands, political agendas, economic situations, and societal expectations drive national-level assessment policies and revisions in England.

This has led to numerous policy reforms and upheaval. Questions remain about whether the current system could be further improved.

This review of research ultimately aims to inform discussions about policy development, improvement, and reform. To identify good – and possibly bad – practice when it comes to assessment policy, we will take an international approach. Looking within and across 10 jurisdictions, we will examine whether assessment systems achieve their stated purposes and what impact they have on student outcomes (namely attainment and motivation to learn).

The commissioning context and broader project

The organisation edpol.net (now known as Centre for Education Systems, CES) commissioned this review through the Charitable Foundation for Education Development. edpol.net – and now the CES - sought to present evidence-based insight into education policy making problems and how to improve them. CES has identified a need to develop an independent education policy knowledge centre, intended ultimately as a knowledge repository to inform education policy in England. The centre aims to provide a comprehensive picture of education policy across a range of international jurisdictions, and to cover key policy topics such as assessment, accountability, school structures, admissions, teacher development and support, curriculum etc.

The broader project involved three phases, namely:

- *A scoping study* - by the National Foundation for Educational Research (NFER). Started November 2022 and completed January 2023 (see Appendix A).
- *An evidence synthesis* - by the UCL EPPI-Centre. Started February 2023 and completed July 2023. The policy comparison table was conducted in October – November 2023. This includes:
- *Ten individual jurisdiction reports*. Available on the EPPI Centre and CES websites.
- A cross-jurisdiction report of research evidence – the focus of this document.
- *A feasibility report* - by the UCL EPPI-Centre. Started February 2023 and completed July 2023. Available by request from the first author of this report.
-

The remainder of this document relates only to Phase 2, unless where otherwise stated. The overarching research questions for Phase 2² are:

- RQ1: What is the current summative assessment system(s) for pupils in general lower and upper secondary education (excluding HE/FE)? Specific aspects to explore: at what ages/stages does the assessment take place; what forms do assessments take (e.g. teacher assessment, examination, coursework); mode of assessment (paper-based, electronic); is it standardised; and how long has the system been in place/when was it introduced?
- RQ2: What are the stated purposes of each summative assessment, i.e. for what explicit purposes are the outcomes of summative assessments used (e.g. grading and sorting; certification; institutional accountability), and are results published?
- RQ3: What is the process through which national-level assessment policy is made? Who specifies what needs to be assessed and how is this determined? (e.g. Is there a regulatory authority or an awarding body specifying minimum content?)
- RQ4: Through what mechanisms is the current summative assessment system evaluated?
- RQ5: What contextual factors are reported to have driven national-level assessment policy and / or any reforms over the last 10 years, and what does this mean for transferability to the context in England?
- RQ6: What does academic research tell us about the extent to which the assessment systems meet the purposes stated in RQ2? What does rigorous, quantitative evidence tell us about the impact of assessment systems and policy on students' motivation to learn, and attainment?

Within **Phase 2**, there are two components: individual jurisdiction reports and a cross-jurisdiction report.

² Phase 1 also aimed to address RQ1-5 through brief mapping of the evidence rather than through detailed mapping and synthesis of the evidence. Phase 3 (the feasibility report) has its own set of research questions, with the aim of considering the potential for adopting a similar approach in other education policy domains (e.g., curriculum, accountability).

The **individual jurisdiction reports** contain a summary of the policy sources for a given jurisdiction, organised by research questions 1-5. The jurisdiction report also includes a summary of any studies found in the research review for that particular research question, in an attempt to address RQ6. These are designed to give an overview of the current policy and research on summative education at the secondary level within a particular jurisdiction. The methods for the policy sources are provided in Appendix B. These individual jurisdiction reports are published separately and are available for the following jurisdictions:

- England
- Estonia
- Finland
- Hong Kong
- Japan
- New Zealand
- Ontario (Canada)
- Poland
- Republic of Ireland
- The Netherlands

This **cross-jurisdiction report** aims to explore summative assessment policy in general secondary education *across* the abovementioned jurisdictions. It mainly draws on a review of published research (see methods in the following sections) and focuses on research question 6. There is, however, a section that considers the similarities and differences between the themes identified in the research and policy sources.

This cross-jurisdiction report takes a holistic look (i.e., over the whole set of research sources) and consists of observations about the generalisability and transferability of findings, and limitations in the evidence base. We had initially envisaged making comparisons between jurisdictions, but heterogeneity and gaps in the evidence makes direct comparisons challenging. Instead, we consider themes that emerged throughout the dataset and how those themes might inform policy development, plus a section on transferability to the English context.

Methods for the research evidence review

Research question

The focus of this report is to address research question 6:

What does academic research tell us about the extent to which the assessment systems meet the purposes stated in the primary studies? What does rigorous evidence tell us about the impact of assessment systems and policy on students' motivation to learn and attainment?

Note that this was a slight modification of research question 6 as originally stated, by removing the restriction to quantitative evidence. This allowed for the inclusion of other research designs, such as interviews and focus groups, to gain greater insights into the barriers and facilitators to effective assessment systems. This expansion of scope is consistent with the inclusion criteria applied by NFER in Phase 1.

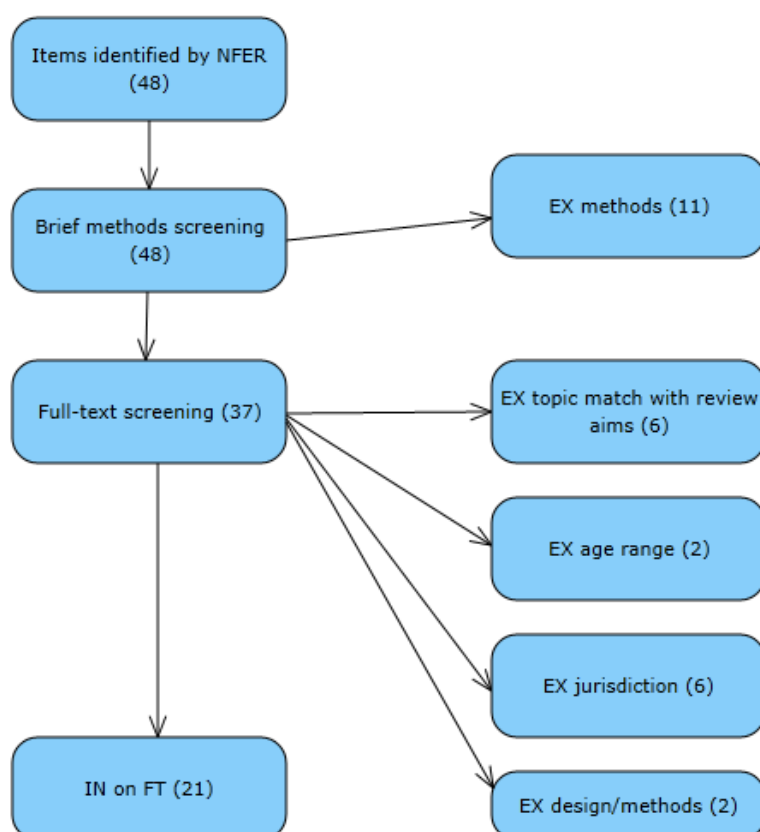
Design

This is a rapid systematic review of the research evidence, with three types of synthesis: thematic summary, thematic synthesis, and a comparative synthesis. For the most part, we followed the processes of a full systematic review, including double screening and double data extraction. The rapid elements were relatively minor: the lack of citation chasing of the included studies and the use of a light touch critical appraisal tool.

Study identification

We used evidence identified from searches undertaken by the National Foundation for Educational Research (NFER). NFER were commissioned at a prior stage to search for documents that would enable a policy and research review to be undertaken. The methodology for this is reported elsewhere³. It was decided at the initiation meeting with the funder for this part of the study, that additional searching would be out of scope. This meant that the evidence synthesised in this report was restricted to the results produced by NFER in the first part of this study.

We only considered evidence identified by NFER from their searching of key databases. At the point we began work on this study, we received 48 research studies included on title and abstract. Figure 1 shows the flow of literature through the screening stages, which are described in detail below.



³ Sharp et al. (ibid).

Figure 1. Flow of literature through the screening process. Note: FT = full-text document, EX = excluded document, NFER = National Foundation for Educational Research. Numbers in brackets represent the number of items (documents) included or excluded for the reasons given. Not all exclusion codes were used in this review, likely because studies were screened for date, language, and publication type at earlier stages in the review.

We first checked the studies were in scope in terms of methodology. Two researchers coded the methods⁴ to a) identify the methods used in each study and b) to rule out any studies where methods were unclear and/or opinion based. After double checking each other's coding, 31 studies were retained at this stage.

Next, we screened the studies on full text using the following inclusion criteria:

- **Publication date**
 - Include if published 2012 or later. Exclude if published 2011 or earlier.
- **Outcomes**
 - Include reports on students' motivation to learn and academic attainment/achievement. Exclude if measured outcomes do not include student motivation or attainment.
- **Population**
 - Include reports on general lower and upper secondary education (ages 14-18/19). Exclude reports on early years, primary/elementary school, further education, and higher education.
- **Intervention**
 - Include reports on assessment systems and policy made at the jurisdiction level (i.e., national or state), relevant to the main summative assessment pathways in general secondary education (academic and/or technical). Exclude reports on localised (e.g., local education authority or school-based) assessment policies.
- **Jurisdictions**
 - Include reports conducted in the following jurisdictions:
 - England (reference jurisdiction)
 - Estonia (PISA high performer; northern European jurisdiction; availability of English language information may be variable)
 - Finland (PISA high performer; Scandinavian jurisdiction)
 - Hong Kong (Asian PISA high performer)¹
 - Ireland (performed similarly to the UK in PISA 2018; European neighbour)
 - Japan (Asian PISA high performer; availability of English language information may be variable)
 - New Zealand (performed similarly to the UK in PISA 2018; Pacific Rim country)
 - Ontario (Canada) (Canada is a PISA high performer; north American jurisdiction; assessment / education systems vary across the

⁴ We coded the type of study methods used (e.g., interviews, cross-sectional surveys) and excluded any that did not meet our inclusion criterion. That is, we only included large-scale primary research studies (qualitative and quantitative), and secondary analysis and systematic literature reviews/ cross-jurisdiction mapping studies. We exclude small-scale quantitative and qualitative studies (operationalised by NFER in Phase 1), non-systematic literature reviews, and non-empirical papers (e.g., think pieces, commentaries, policy documents).

- Canadian provinces so a focus on a single province / jurisdiction is helpful)
 - Poland (PISA high performer; northern European jurisdiction; availability of English language information may be variable)
 - Netherlands (performed similarly to the UK in PISA 2018; northern European jurisdiction; availability of English language information may be variable).
 - Exclude if the report relates to other countries or jurisdictions.
- **Study type**
 - Large-scale primary research studies (qualitative and quantitative), and secondary analysis and systematic literature reviews/ cross-jurisdiction mapping studies. Exclude small-scale quantitative and qualitative studies (operationalised by NFER in Phase 1⁵), non-systematic literature reviews, and non-empirical papers (e.g., think pieces, commentaries, policy documents).
- **Publication type**
 - Include peer-reviewed and grey literature. Exclude if the item is a policy document or an opinion piece.
- **Publication language**
 - English language only. Exclude if published in languages other than English.

Two researchers did this screening and double checked all results. At this stage, 16 studies were excluded, leaving 21 in the review.

Data Extraction

We developed and piloted a data extraction code set for use on the included studies. Some of the items in the data extraction tool were developed a priori, while others were developed inductively. This means that they were developed (added or amended) as we interacted with the studies and new themes emerged. This allowed us to capture information that would be relevant to the review that was not anticipated. For example, we had not anticipated the difference between purposes and uses of the assessment; this distinction in the code set was added after piloting suggested there was a distinction in the evidence base.

This coding aimed to provide us with suitable evidence to help us address research question six, but also reflected elements of research questions one to five (e.g., characteristics of the assessment system). The data extraction code set included extracting information about:

- **The assessment characteristics** – a description of the summative assessment, a description of the summative assessment system evaluated in the paper, the policy's target age groups, the form of assessment (e.g., teacher assessment, exam assessment, coursework assessment, etc), the mode of assessment (e.g., paper, electronic, etc), and whether the assessment was standardised.
- **The stated purposes of the assessment**, e.g., to determine pupil access to further and higher education.
- **The stated uses of the assessment** (could be same as purpose but also different).
- **The study characteristics** – the year the study commenced /publication year, the jurisdiction(s), the sampling for the evaluation, key aspects about the sample, such as age of participants, and sample size.

⁵ Small-scale quantitative or qualitative studies were those with fewer than 50 cases.

- **Results of the study/evaluation.** We extracted key information reported by the authors that addressed the paper aims and in particular covered findings on attainment, stress and anxiety reported as related to the evaluated summative assessment, and any findings reported that related to students' motivation to learn.
- **Debates about assessment.** In discussions with the funder, we identified that there were some key themes of interest to extract if any of the papers reported on them. These included, 'high stakes versus low stakes assessment', 'potential harms of summative assessment to pupils' and 'reported implications of summative assessment for future policy making'.

The full, final data extraction tool can be found in Appendix C. The data extraction was undertaken by three researchers. Each study was extracted by one of the three reviewers and then checked by a second reviewer to ensure consistency and accuracy of coding. Some discrepancies were discussed amongst all three reviewers, when a third opinion was deemed necessary.

Critical Appraisal

We critically appraised the included studies, but it was challenging. This was largely because the scope of the review was wide. For example, we included a wide range of assessments within the permitted age range for the review, with varied purposes. The studies were also drawn from a range of jurisdictions that had disparate cultural, political, and policy contexts, and a range of methodologies (qualitative, quantitative, and mixed method), and a range of publication types (journal, reports, online sources). This level of heterogeneity made it difficult to attempt to apply a standard systematic review critical appraisal tool.

Instead, we opted for a more pragmatic approach that screened for: author's claims about their study's overall quality, any reported methodological limitations that may call into question the validity and reliability of analysis conducted, and any ethical concerns raised by the authors. We also noted any concerns from the coders (our research team) about the research and analysis process, such as a lack of clarity of methods.

The critical appraisal judgements were undertaken by three researchers, and we double coded and checked each other's entries. This was a quality check of our own processes. The results of the critical appraisal are shown in Table 1.

Synthesis

The synthesis occurred in three main stages: thematic summaries, thematic synthesis, and a comparative synthesis that ties in the findings from the policy documents.

The **thematic summaries** produce a descriptive account of the evidence (Thomas, O'Mara-Eves, Harden, Newman, 2017)⁶. We started by generating tables that showed the characteristics and frequencies of the studies in the review. These tables, and their accompanying descriptive text, give an overview of the evidence base. The tables and text within the thematic summaries are organised in the same way as the themes in the data extraction tool, i.e., grouped by assessment characteristics, study characteristics, critical appraisal, debates in the field, and study findings.

Thematic synthesis, in contrast, aims to knit the findings together in a way that creates new understandings. Rather than simply just describing the state of the evidence as in thematic

⁶ Thomas, J, O'Mara-Eves, A, Harden, A, & Newman, M. (2017). Synthesis methods for combining and configuring textual or mixed methods data. In D Gough, S Oliver, and J Thomas (eds.), *An introduction to systematic reviews* (2nd ed.), London: SAGE.

summaries, the synthesis attempts to develop new theories or frameworks by exploring points of convergence and difference amongst the studies (Thomas et al., 2017). We try to move beyond the labels and frequencies of the thematic summary to consider the content of each theme, such as the magnitude and direction of an effect, or participants' views about a particular phenomenon.

We started by generating tables of the line-by-line coded text associated with each theme. We were then able to see how the studies dealt with a particular concept. For example, we looked at all the data that had been extracted for the theme of attainment findings to see what the studies could tell us about the relationship between assessment systems and attainment. We highlight points of similarity and divergence. In practise, for this review, synthesis was impeded for some of the themes because of the vast heterogeneity in the data set. In other words, there was more divergence than similarity. At the end of the thematic synthesis, we generated a conceptual framework that attempted to show how the different themes might relate to each other.

The **comparative synthesis** attempts to bring together our understanding of summative assessment that we have gained through exploring the research evidence (the focus of this report) and what we have learned through examining the policy sources (the focus of the individual jurisdiction reports; methods available in Appendix B). The approach taken here is most similar to a 'segregated mixed method synthesis' as detailed by Sandelowski, Voils, and Barroso (2006)⁷, in which two different evidence types are synthesised separately (in this case, the research studies and the policy sources), and then brought together in a mixed research synthesis in a process that they called configuration. While we formally synthesised the research evidence, we did not undertake a formal synthesis across the policy sources, so this report cannot truly be called a mixed methods synthesis. Instead, we acknowledge the procedure of comparing the themes identified across the two evidence types (research papers and policy sources) by using the label 'comparative synthesis'.

For the comparative synthesis, we took as a starting point the conceptual framework that was generated through the thematic synthesis and then compared that with core themes that were identified across the policy sources. We were particularly interested in how well the two types of evidence overlapped and which aspects were unique. That is, we looked for points of convergence and divergence.

⁷ Sandelowski M, Voils CI, Barroso J. (2006). Defining and designing mixed research synthesis studies. *Res Sch.* 13(1):29.

Table 1. Summary of critical appraisal of the included papers

Critical appraisal questions	Number of studies	Notes
Did the research clearly state they had ethical approval or an advisory board?		This question was designed to capture whether the study had some sort of external accountability, ethical checks, or quality assurance
Yes	2	Ethics approval or an advisory board were clearly mentioned
No	19	No explicit mention of ethics approval or an advisory board
What was the richness of the study (amount of content that speaks to the review's questions)?		This question was designed to indicate how well the study could inform our review's aims. This is akin to the relevance of the study to the review's focus.
High	13	Rich coverage of our review aims
Low	8	Limited scope related to our review aims
Are the methods transparent and clear?		This question was intended to identify studies without adequate reporting of their methods
Yes	18	Good transparency and clarity of methods
No	3	Three studies had unclear methods so that replicability would be difficult and/or aspects of their designs were not transparent
Are there any serious limitations reported by the authors?		This question was designed to identify any major concerns about the study as identified by the primary study authors
Yes	2	Limitations had to be explicitly mentioned by the primary study authors. This does not mean other studies did not have serious limitations. Baird 2016 reported that a critical limitation of their study was that "students prior attainment is not controlled for in associations between predictability views and examination results" (P. 29). They also mention sample bias towards higher achieving students which could have affected the results. The other study was Baird 2019. Some key serious limitations were mentioned as follows: representativeness of the interview and quant data (e.g., "the data cannot be said, even closely, to represent a randomised control trial – P.77), modular and linear assessment routes were not available in each subject in every academic year which

Critical appraisal questions	Number of studies	Notes
		meant outcomes for these might be confounded with the standard setting process, the analysis was limited by missing values (especially of SES students), different teaching approaches were noted by the authors between modular and linear assessments which could affect the results, and the assessment route is almost entirely confounded with school of entry.
No	19	Either limitations noted but not considered serious by the authors, or Limitations not adequately discussed
Does the reviewer have any concerns about the study?		This question was designed to identify any major concerns about the study as identified by the systematic reviewers
Yes	17	In summary, the key issues were: • Authors acknowledging limitations but failing to deal with them/mitigate them. • Authors failing to acknowledge some limitations/ Insufficient discussion of limitations. • Limited search strategy and limited discussion of their quality assurance processes (this was a review). • Poor reporting of methods including key details of things like the sample. • Lack of clarity about the groups being compared. • Lack of sample representativeness • Concerns about author impartiality. • Concerns about missing data. • Seemingly weak rationale for approach and decisions about what to analyse. • Lack of clarity in presentation of results and merging of key topics. • Small sample size.
No	4	Reviewers felt no comments on serious limitations were necessary.

Methods for the cross-jurisdictional comparison table

The aim of the comparative tables was to provide a descriptive analysis ‘at a glance’ of the key policy differences and similarities between the jurisdictions.

The original review of policy sources took place between February and June 2023. It is important to note that this review reflects the evidence we found at that time. No updating has been done between the project completion and the publication date.

The approach taken for the comparative table was to use the existing individual jurisdiction briefs as our data sources. We extracted data in line with the research questions as specified in the Methods of this report and specific features of interest identified by the funder. We supplemented data extracted from the briefs with searches using desktop AI from BING and ChatGPT technology. These AI checks were conducted in November 2023. In all cases, we verified the response by the AI by following links to the cited documents or identifying the relevant documents through a follow-on search in Google. We only extracted information for each research question as stated in these sources and did not apply any interpretation. Between the briefs and AI, we managed to address all questions in the comparison table. A final step was for the two researchers to check each other’s entries, and to ask an independent subject expert to review the content and presentation of this part of the review work.

The tables were organised by research question. The table rows were determined in conjunction with the funder, and the columns represent each jurisdiction. While we provided a textual summary to accompany the tables (see below), we did not draw any conclusions about whether any jurisdiction—or its components—are ‘better’ or ‘worse’ than any others. We acknowledge that this may have limited value in terms of identifying ways in which a given system could be improved; it may simply indicate that there are alternative options.

The comparison table was completed after the individual jurisdiction briefs were finalised. There was a gap of 4 months between the two. The table is designed as a snapshot of key features of interest rather than being a summary of what is in the text of the brief, and therefore some new information may be introduced as new in the table. It should be seen as complementary to the individual reports, rather than a summary of them.

Results of the thematic summary

The review included 21 primary research studies from seven jurisdictions (Table 2). The jurisdictions ranged in the number of studies provided for the review. Most of the studies were from England ($n = 12$), followed by Ireland ($n = 7$); there were no studies included for Estonia, Poland, and Japan.

Table 2. Number of studies for each jurisdiction of interest

Jurisdiction	Number of studies (n = 21)
England	12
Estonia	0
Finland	2
Hong Kong	2
Japan	0
New Zealand	2
Ontario (Canada)	1
Poland	0
The Netherlands	1
The Republic of Ireland	7

Note. Some studies reported on more than one jurisdiction, so the sum of column two is greater than the total number of studies in the review (n = 21).

The key characteristics of the summative assessment systems are reported in Table 3. Examples of systems include the General Certificate of Secondary Education (GCSE), the Higher Irish Leaving Certificate and The English Baccalaureate. The qualifications generally covered age ranges of either 14-16- or 17–18-year-olds. Most of the systems included in this review were what would be termed ‘high stakes’ assessment. A high-stakes test is a test with important consequences for the test taker, and which has important benefits, such as a qualification needed for further education or for a profession/employment.

Most of these assessments (n = 15) used examinations to assess the required standard. Examinations were either used as a standalone method or in combination with teacher assessments, coursework or some sort of practical investigation for the student to perform. From the primary studies, it was not always clear what the medium of the assessment was used, although paper-based assessments were explicitly mentioned in some studies. Some studies mentioned standardisation as a feature of the system, but few specified this detail.

Table 3. Reported Characteristics of the Summative Assessment Systems

Study	Jurisdiction of Interest	Name of Summative Assessment System(s)	Targeted Age Group and/or Stage of Education	Format of Assessment
Baird (2016)	Republic of Ireland	Higher Irish Leaving Certificate	Median age of examinees = 18	• Examination • Practical Investigation
Baird (2019)	England	General Certificate of Secondary Education (GCSE)	14-16	• Teacher assessment • Examination • Coursework
Banks (2015)	Republic of Ireland	Leaving Certificate (LC)	Upper Secondary School	Examination
Benton (2014)	England	Entry-Level Qualifications (ELQs)	• 16 • Final year of compulsory schooling	Unclear
Brown (2022)	England	General Certificate of Secondary Education (GCSE)	16	• Examination • Coursework
Cadwallader (2013)	England	The English Baccalaureate (EBacc)	Students taking their GCSEs	Unclear
Carroll (2018)	England	• Programme for International Student Assessment (PISA) • General Certificate of Secondary Education (GCSE)	Unclear	Unclear
Elwood (2017)	Republic of Ireland	Leaving Certificate (LC)	• 16-18 • Upper secondary students	Examination
Hargreaves (2020)	Ontario, Canada	Mid-stakes testing (EQAO)	• Grade 9 • Secondary school	Unclear

Study	Jurisdiction of Interest	Name of Summative Assessment System(s)	Targeted Age Group and/or Stage of Education	Format of Assessment
Hepworth (2021)	England, Republic of Ireland	Programme for International Student Assessment (PISA)	15	Unclear
Jerrim (2022a) (Test anxiety)	England	General Certificate of Secondary Education (GCSE)	Year 11	Examination
Jerrim (2022b) (Power)	England	General Certificate of Secondary Education (GCSE)	Year 11	Examination
Lessof (2018)	England	General Certificate of Secondary Education (GCSE)	• Key Stage 4 (KS4) • Year 11.	Examination
Machin (2018)	England	General Certificate of Secondary Education (or GCSEs)	16	• Teacher assessment • Examination • Coursework
McCoy (2018)	Republic of Ireland	Leaving Certificate (LC)	Students in their sixth year	Examination
Nieminen (2022)	Finland	Assessment for Learning (AfL)	• 12 (6th grade) • 15 (9th grade)	• Teacher assessment • Examination • Coursework • Portfolio, product (e.g. a computer program), performance (e.g. drama, music), etc.
Ofqual (2012)	England	Advanced Level General Certificates of Education (A-levels)	18	Examination
	Finland, Hong Kong, Netherlands, New Zealand, Republic of Ireland	Varied, depending on jurisdiction and subject	Final stage of secondary school for the given jurisdiction	Varied, depending on jurisdiction and subject.

Study	Jurisdiction of Interest	Name of Summative Assessment System(s)	Targeted Age Group and/or Stage of Education	Format of Assessment
Putwain (2016)	England	General Certificate of Secondary Education (GCSE)	Students in their final year of secondary school	Examination
Smyth (2012)	Republic of Ireland	Junior Certificate Leaving Certificate	15 18	Examination
Tong (2015)	Hong Kong	School-based Assessment (SBA), an integral part of the English Language assessment for the Hong Kong Diploma of Secondary Education (HKDSE)	• 15 • Secondary 5 (i.e., Grade 11)	• Classroom-based assessment • Provision and receipt of feedback
Zhao (2022)	New Zealand	National Certificate of Educational Achievement (NCEA)	Years 11-13	• Teacher assessment • Examination

We extracted data on the study design and sample characteristics (see Table 4). Many of the included studies (18/21) had nationwide coverage; three focused on specific regions within a jurisdiction. The sample size varied widely, from 35 participants to 675,000 participants. The study participants were mostly students (18/21) but also included teachers (3 studies), educational/policy specialists (1 study) and guidance counsellors (1 study) and parents (1 study). While most of the studies used primary data collection, a minority used other methods such as secondary analysis of large-scale datasets (7/21 studies).

In addition to the study characteristics, we also extracted information reported within the studies about the purposes and uses of the assessments (Table 5). In some of the studies, such as Baird (2019), the purposes and uses overlapped, and in others, such as Carroll (2018), the stated purposes and uses differed. For example, in the Carroll (2018) study, the reported *uses* of the assessment, the GCSE, was to determine which students are eligible for higher education and to allow employers to differentiate between candidates. In contrast, the reported *purposes* of these assessments in Carroll (2018) were to progress students to higher education, to assess whether learning objectives have been achieved, or assess student progress (i.e., pupil knowledge, skill acquisition), and to enable international comparisons to be made.

Our final table (Table 6) relates to the findings of the individual studies. The majority of the included studies reported findings that related to the stated purpose of the assessment system (17/21). Other outcomes of interest to this review that were reported were attainment (13/21), stress or anxiety (11/21), and motivation to learn (6/21).

Table 4. Study Design and Population Sample of the Studies

Study	Publication Type	Data collection method	Population Sample				
			Coverage	Characteristics	Actors within the education system	Stage of Education and/or age	Size
Baird (2016)	Journal Article	Questionnaire	Nationwide	Universal (General population)	Students	Higher Level Leaving Certificate (LC) examination sitters	1002
Baird (2019)	Research Report	Semi-structured Interview	Nationwide	Universal (General population)	Teachers	Not applicable	35
Banks (2015)	Journal Article	• Survey • Interview	Nationwide	Universal (General population)	Students	Final year of secondary school (sixth year)	• Student surveys (N = 748) • 6-person group interviews (N = 53 groups)
Benton (2014)	Journal Article	Longitudinal Study of Young People in England (LSYPE) Data	Nationwide	Universal (General population)	Students	13-16-year-olds	Approximately 10,000
Brown (2022)	Journal Article	Systematic Review	Varied according to study	Universal (General population)	Students	GCSE sitters	22 included articles
Cadwallader (2013)	Research Report	GCSE and A-Levels results	Nationwide	Universal (General population)	Students	16-year-old GCSE sitters	Approximately 675,000
Carroll (2018)	Research Report	National Pupil Database (NPD)	Nationwide	Universal (General population)	Students	Unclear	5,194

Population Sample							
Study	Publication Type	Data collection method	Coverage	Characteristics	Actors within the education system	Stage of Education and/or age	Size
Elwood (2017)	Journal Article	Interview	Nationwide	Universal (General population)	Students	Unclear	81
Hargreaves (2020)	Journal Article	Interview	Region: Ontario	Unclear	Teachers	Not applicable	Not clearly specified as the paper reports on several research projects
Hepworth (2021) (Section 3.2)	Research Report	Interview	Nationwide	Universal (General population)	Educationalists and policy experts	Not applicable	Unclear
Jerrim (2022a) (Test anxiety)	Journal Article	Programme for International Student Assessment (PISA) data	Nationwide	Universal (General population)	Students	Year 11 (15-year-olds)	4,914
Jerrim (2022b) (Power)	Journal Article	Programme for International Student Assessment (PISA) data	Nationwide	Universal (General population)	Students	15/16-year-olds	Approximately 4800
Lessof (2018)	Research Report	Longitudinal Study of Young People in England (LSYPE) data	Nationwide	Universal (General population)	Students	Year 11 completers	Each chapter has a different total sample size. Chapter 2: 9,350 Chapter 3: 4,200 Chapter 4: 4,091
Machin (2018)	Research Discussion Paper	Administrative data on the census of students in state schools	Nationwide	Universal (General population)	Students	16-year-olds	383,730

Population Sample							
Study	Publication Type	Data collection method	Coverage	Characteristics	Actors within the education system	Stage of Education and/or age	Size
McCoy (2018)	Research Report	State Examinations Commission Data	Nationwide	Universal (General population)	• Students • School principals • Guidance counsellors • Parents	Sixth year (students)	• State Examinations Commission Data pertaining to sixth year students (N = 153,156) • Number of interviewees is unclear
Nieminen (2022)	Journal Article	Survey	Nationwide	Universal (General population)	Students	Ninth graders (15-year-olds)	278
Ofqual (2012)	Research Report	Review of different jurisdictions	Nationwide and regions	Unclear	Not applicable	Not applicable	Not applicable
Putwain (2016)	Journal Article	Questionnaire	Region: North-west of England	Universal (General population)	Students	Final year of secondary education (Year 11)	325 school students (male n = 142, female n = 183)
Smyth (2012)	Journal Article	• Questionnaire • interview	Nationwide	Universal (General population)	Students	Third year	• Questionnaire (N = 897) • 6-Person Group Interview (N = 47)

Study	Publication Type	Data collection method	Population Sample				
			Coverage	Characteristics	Actors within the education system	Stage of Education and/or age	Size
					Students	Sixth year	• Questionnaire (N = 748) • 6-Person Group Interview (N = 53)
Tong (2015)	Journal Article	• Survey • interview	Nationwide	Universal (General population)	Students	Secondary 5	• Surveys (N = 423) • Interviews (N = 45)
Zhao (2022)	Journal Article	Experimental research design	Nationwide	Universal (General population)	Students	Secondary School, years 10-13 (16-18 years of age)	479

Table 5. Reported Purposes and Uses of Summative Assessment Systems

Study/ stated purpose	To determine pupil access to tertiary education		To determine access to vocational or technical education, internships		To progress to higher levels of secondary school education		To assess whether learning objectives have been achieved		To assess school- or system-level performance for quality assurance or accountability		To assess school-level performance to inform parents/careers decisions about school applications		To enable international comparisons		To incentivise teachers and/or school managers		To allow employers to differentiate between candidates		Others	
	P	U	P	U	P	U	P	U	P	U	P	U	P	U	P	U	P	U	P	U
Baird (2016)	✓				✓													✓		
Baird (2019)					✓	✓			✓	✓										
Banks (2015)	✓	✓																✓		
Benton (2014)						✓	✓												1	
Brown (2022)		✓							✓	✓						✓		✓	2	
Cadwallader (2013)										✓									3	
Carroll (2018)		✓			✓		✓						✓					✓		
Elwood (2017)	✓	✓					✓													
Hargreaves (2020)										✓	✓	✓								
Hepworth (2021)								✓					✓							4

Study/ stated purpose	To determin e pupil access to tertiary educatio n		To determin e access to vocationa l or technical education , internshi ps		To progress to higher levels of seconda ry school educatio n		To assess whether learning objectiv es have been achieved		To assess school- or system- level performanc e for quality assurance or accountabili ty		To assess school-level performanc e to inform parents/care rs decisions about school applications		To enable internation al compariso ns		To incentivi se teachers and/or school manager s		To allow employers to differentia te between candidate s		Other s	
Jerrim (2022a) (Test anxiety)		✓				✓	✓											✓		
Jerrim (2022b) (Power)										✓										
Lessof (2018)		✓				✓												✓		
Machin (2018)	✓	✓	✓							✓						✓		✓		
McCoy (2018)	✓	✓																✓	5	
Nieminen (2022)							✓	✓											6	6
Ofqual (2012)	✓	✓					✓	✓												
Putwain (2016)		✓																✓		
Smyth (2012)		✓				✓												✓		
Tong (2015)							✓	✓				✓								
Zhao		✓				✓	✓		✓									✓		

Study/ stated purpose (2022)	To determin e pupil access to tertiary educatio n	To determin e access to vocationa l or technical education , internshi ps	To progress to higher levels of seconda ry school educatio n	To assess whether learning objectiv es have been achieved	To assess school- or system- level performanc e for quality assurance or accountabili ty	To assess school-level performanc e to inform parents/care rs decisions about school applications	To enable internation al compariso ns	To incentivi se teachers and/or school manager s	To allow employers to differentia te between candidate s	Other s

Note. P = purpose, U = use. 1 = To offer an alternative for, and engage, certain groups of learners (e.g. students with prior records of low attainment or disabilities). 2 = To ensure equality within education and the demonstrability of a wide range of skills and abilities. 3 = To ensure equality within education and increase the student uptake of particular subjects. 4 = To help teachers develop their pedagogical practice, thereby reducing the need for later intervention. 5 = To ensure fair grading and to encourage the uptake of certain subjects. 6 = To create the best learning opportunities for every learner.

Table 6. Reporting of outcome types in the primary studies

Purpose			Did the studies report findings relating to the following outcomes:		
Study	Were the purpose(s) of the featured summative assessment system explicitly reported in the study?	Did the study evaluate the purpose(s) reported in the study?	Attainment	Motivation to learn	Stress/anxiety
Baird (2016)	Yes	Yes	Yes	No	No
Baird (2019)	Yes	Yes	Yes	Yes	Yes
Banks (2015)	Yes	No	No	No	Yes
Benton (2014)	Yes	Yes	Yes	No	No
Brown (2022)	Yes	Yes	Yes	Yes	Yes

Cadwallader (2013)	Yes	Yes	Yes	No	No
Carroll (2018)	Yes	Yes	Yes	No	No
Elwood (2017)	Yes	Yes	No	No	No
Hargreaves (2020)	Yes	Yes	No	No	Yes
Hepworth (2021)	Yes	Yes	Yes	Yes	No
Jerrim (2022a) (Test anxiety)	Yes	No	Yes	No	Yes
Jerrim (2022b) (Power)	No	No	Yes	No	No
Lessof (2018)	No	No	Yes	No	Yes
Machin (2018)	Yes	Yes	Yes	No	No
McCoy (2018)	Yes	Yes	Yes	Yes	Yes
Nieminen (2022)	Yes	Yes	No	No	Yes
Ofqual (2012)	Yes	Yes	No	No	No
Putwain (2016)	No	No	Yes	No	Yes
Smyth (2012)	No	No	No	Yes	Yes
Tong (2015)	Yes	Yes	No	No	No
Zhao (2022)	Yes	Yes	No	Yes	Yes
Total “Yes”	17	15	13	6	11

Key conclusions from the thematic summary

- The review included 21 primary research studies from seven jurisdictions. Most of the studies were from England (n = 12). This was followed by the Republic of Ireland (n = 7); Finland, Hong Kong, and New Zealand (n = 2 each); and Ontario and The Netherlands (n = 1 each). There were no eligible studies identified for Estonia, Poland, and Japan.
- The most reported *purpose* of the assessment system was to assess whether learning objectives have been achieved (n = 9), then to enable progression through levels of education (n = 8).
- The most reported *use* of the assessment system was to enable progression through levels of education (n = 14), with the majority of those focused on progression to higher education (n = 12). The next most common use was to allow employers to differentiate between candidates (n = 11).
- Most of the studies explicitly stated that the assessment used examinations to assess the required standard (n = 15), sometimes in combination with other assessment types such as coursework.

Results of the thematic synthesis

We synthesised five key themes related to what the studies identified as ‘purposes’ of the assessment within each paper. These themes were: Progression, Student learning, Quality Assurance, International Comparisons, and Fairness and Equality. We also synthesised three key outcomes of the assessments: Attainment, Motivation, and Stress/Anxiety. The key findings for each are summarised below. In all cases, the reader is reminded that most of the data came from England (n = 12) or Ireland (n = 7), and so the findings may not fully represent the range of possible themes that might be evident in other contexts and cultural settings.

Themes about purposes of the assessment

We identified five broad themes that studies mentioned as the key purpose of the assessment policy or system. These were: to determine progression to the next stage of education; to assess student learning; to enable quality checks of schools or the education system; to enable international comparisons; or to establish fair and equitable assessment. Many of the studies mentioned more than one of these themes, which we elaborate on below.

Progression to the next stage of education

This theme was related to assessments which allow students to progress beyond their current level of education. These included:

- Assessments that enable students to access higher levels of secondary school education. For example, in England, this usually involves pupils taking GCSEs at age 16 to progress to A-levels at age 18.
- Assessments that determine pupil access to tertiary/university/higher education. For example, A levels in England are used by pupils to access higher education in universities.
- Assessments that enable students to progress to vocational or technical education.

The debates in the included papers usually discussed progression using what is termed ‘high stakes’ assessment, whereby progression hinges on passing qualifications at a certain

threshold of grade to progress to the next level. Additionally, students are noted to progress linearly (from one level of qualification to another).

Eight studies were coded in this theme and had 'progression' of pupils as a stated purpose of the assessment. Of these eight studies, five provided findings from the study about this theme. The results are presented in Appendix D.

Overall, the studies indicate that the assessment systems do achieve the purpose of allowing progression to the next stage of education, whether it be upper secondary, tertiary, or vocational education. In particular, the studies on the GCSE assessments for 14-16 in England reported that it was an appropriate pathway into the next stage of education (Baird 2019; Ofqual 2012). There were, however, some concerns across the different assessments about whether the assessment systems equipped the students with the learning they needed in the future (e.g., Baird 2016). Relatedly, the importance of learning practical skills as part of the assessment process was raised (Ofqual, 2012). Further, because of the emphasis on progression to the next stage of education, students may place a strong emphasis on maximising grades (Elwood, 2017). Finally, where progression to further education and employment hinges on very strict threshold assessment levels, failure to meet the threshold could impact students' access to education and employment (Machin 2018).

Student learning

This theme covered findings from the included studies about whether stated pupil learning objectives had been fulfilled, and/or findings about knowledge or skill acquisition, and what aspects of the assessments facilitated or hindered this. This theme also covered how well the assessments prepared students for future study. Seven studies were coded in this theme (see Appendix E).

The assessments tried to build in the goal of improving student knowledge as an outcome, but the main aim of the assessments was still viewed by students as leading to a qualification to help them progress to other qualifications (Elwood 2017; Nieminen 2022; Tong 2015). Some studies noted the complexity of assessments that combines formative and summative elements (Nieminen 2022; Tong 2015). Helpful feedback, relevant to the task, was considered important (Tong 2015), as was the provision of content specific to the subject being taught (Ofqual 2012).

In general, GCSEs and A levels were considered to capture individual student learning well (Carroll 2018; Ofqual 2012), but some students complained that certain aspects (coursework and exams) were not explained correctly to students (Brown 2022). A commonality among the above studies, was the need to include student voice and perspectives when designing or revising secondary level assessments as they are best placed to judge what methods work best for them and which also meets appropriate learning outcomes (Brown 2022; Elwood 2017; Nieminen 2022; Tong 2015).

Quality assurance

This theme was related to assessments that were conducted to determine the quality of schools or implemented systems, thereby ensuring transparency and accountability while improving learning and teaching methods. Specifically, governments used such assessments to identify under-performing schools as a basis for intervention and/or school

closure, monitor the performances of teachers (Baird, 2019; Hargreaves, 2020), as well as justify the introduction of new systems or educational policy reforms (Elwood, 2013).

Four studies were coded as having 'quality assurance or accountability of school or system' as a stated purpose, but only one had provided findings about the theme. The one study with data findings (Hargreaves 2020), from Ontario Canada, found that improvements in student attainment scores gave staff confidence in their teaching, especially when teaching students with learning disabilities. The report cautioned against EQAO being used to promote improvements and interventions in particular schools and districts and to rank schools.

As only one study addressed the theme, it was not possible to conduct a synthesis of combined literature, which usually requires findings from two or more studies. Further, it should be noted that this study did not adequately discuss any limitations in methodology. While the author appropriately notes their impartiality on pages 400-402, whether this affected their interpretation of the data is unclear. Even if the themes and findings accurately represent the data, the expression of the findings are strong given uncertainties around generalisability.

Given the above, the lack of studies assessing whether assessment systems meet this stated purpose may suggest a gap in the literature. This conclusion, however, is strictly based on the studies that the findings of this report are grounded in. A more comprehensive search strategy, focused more on quality assurance, may be necessary to check if this is indeed a literature gap.

International comparisons

This theme was related to assessment systems seeking to enable international comparisons between separate countries, specifically the Programme for International Student Assessment (PISA). In facilitating cross-country comparisons, PISA assumes the role of a measure used by countries to attain a deeper understanding of the strengths and weaknesses of their education systems, monitor effective educational trends, and determine ways to improve educational policies over time by learning from other countries (Hepworth, 2021).

Four studies were coded as having assessments with the purpose of 'enabling international comparisons'. Only two of these studies provided some findings about the theme (Carroll 2018 and (Hepworth 2021) and were studies based in England and Republic of Ireland. The search strategy used within this review was not conducted specifically to examine the effectiveness of PISA and a synthesis of the two studies under this theme, which are about PISA, may not fairly represent a collective perspective of relevant and appropriate extant literature on literature about international comparisons. Furthermore, two studies on PISA are not sufficient evidence to make judgements or summary statements about PISA. Further, it should be noted that both studies did not adequately address any limitations in their methodology, which may call into question the validity or reliability of the content.

Fairness and equality

Four papers mentioned establishing fairness and equality as a purpose of the assessment system. While all four studies provided findings about this theme, the aspects of the assessment system addressed in each study were slightly different. The results are presented in Appendix F.

The studies discussed a wide range of ways that fairness and equality issues permeate assessment. Examples of unfairness and concern, include gender stereotyping, teacher bias, subject bias, and inequalities in the grading structures (Benton 2014, Brown 2022, Cadwallader 2013, and McCoy 2018). Additionally, fairness and equality (or perception of it by students), were seen to be linked to low achieving students staying in education and their attainment (Benton 2014). Assessment systems that place greater emphasis on specific subjects to promote a robust education system for all, have been largely critiqued by stakeholders as being unfair to students who might not excel, or be interested in, said subjects (McCoy 2018). Academic subjects were argued to sometimes be prioritised over other 'marginal' subjects, such as art (Cadwallader 2013).

Summary of findings about purposes of the assessment

- Research question 6 asked what the academic research tells us about the extent to which the assessment systems meet the stated purposes. We found mixed evidence, mostly from England and Ireland, on this theme: some purposes are generally met (with varying levels of caveats), while others are under-researched.
- Assessment systems do achieve the purpose of allowing progression to the next stage of education, whether it be upper secondary, tertiary, or vocational education.
- Systems were generally found to capture individual student learning well.
- There was insufficient evidence to determine whether the assessment policies enable quality assurance or auditing of the quality of schools or the education system.
- PISA may be used for international comparisons of education systems in terms of student performance, but whether it adequately achieves this purpose is not answerable in this review due to the study identification methods used.
- The purpose of ensuring fairness and equality was not always met, with concerns raised about gender stereotyping, teacher bias, subject bias, and inequalities in the grading structures.
- Concerns were raised about whether the assessment systems:
 - equipped the students with the learning they needed in the future, particularly practical skills.
 - overemphasised maximising grades at the expense of learning.
 - risked limiting short- to medium-term opportunities if students failed to pass threshold grades.
 - become overly complex when the system combines formative and summative elements.
 - were adequately explained to the students.
 - prioritised certain core academic subjects (such as maths) over other 'marginal' subjects (such as art) in ways that could be unfair to students with skills or interests in the latter.
- Facilitators to improved learning included:
 - providing helpful, relevant feedback
 - providing content specific to the subject being taught
 - including student voice when designing or revising secondary level assessments

Themes about the study outcomes

Research question 6 asked about the impact of assessment systems and policy on students' motivation to learn and attainment, so data was extracted on those themes. We also identified a recurring theme relating to student stress or anxiety, which we additionally extracted and synthesised.

Attainment

This theme was related to the outcome of student educational attainment. This refers to the successfully completion of an assessment for the award of some qualification, or about academic achievement. Successful completion of a level of education refers to the achievement of the learning objectives of that level, typically validated through the assessment of acquired knowledge, skills and competencies. Thirteen studies were coded with the outcome of student attainment, and the results are presented in Appendix G.

Some studies raised concerns about reforms of qualifications on attainment (Baird 2016; Brown 2022; McCoy 2018). For example, student participants raised concerns about the poor linkage between the content of the taught subjects and the summative assessment (Baird 2016; McCoy 2018). However, other studies found no issues with the assessed qualifications – e.g., no evidence of a positive or negative effect of the assessment (Benton 2014) and GCSEs attainment not being found to be associated with worry, fear of failure or test anxiety (Jerrim 2022a, b).

Facilitators of attainment included the support of teachers (Brown 2022), enthusiasm for the subject (Hepworth 2021), student positive perception of examination (Baird 2016), student choice in subjects studied (Cadwallader 2013), doing separate science subjects rather than combined (Carroll 2018), doing homework (Lessof 2018), getting enough sleep (ibid) and school attendance (ibid). The concept of 'academic buoyancy', which describes student ability to withstand and respond positively to setbacks, was cited as a key attribute for overcoming anxiety about GCSE examinations (Putwain 2016). On the other hand, disadvantage and social economic status was cited as a key factor separating students into higher and lower achievement (Baird 2019; Lessof 2018), and success often hinged on tight threshold grading (Machin 2018).

In summary, assessment systems and policy appear to have a relationship with students' attainment, but in complex ways.

Student motivation

This theme was related to student motivation to learn and attitudes about assessment. Six studies were coded with the outcome of student motivation, and the results are presented in Appendix H.

Two studies cited a difference of opinion when considering the impact of assessments on motivation (Baird 2019, Brown 2022). Those teachers who thought assessment and examinations were good for motivation, expressed this view because the assessments were perceived to raise student expectations, and taught them resilience (Baird 2019). Some teachers considered motivation altered according to the context. For example, the purpose of the test: higher levels of motivation for tests which had individual consequences (Zhao 2022), and for learning material they would be assessed on (McCoy 2018).

Motivation was perceived by teachers to less good when students worked in groups or did not have the opportunity to resist or redo results (Baird 2019). Teachers were cited as important factors supporting student motivation, through their own enthusiasm of the

subjects and support with revision (Brown 2022, Hepworth 2021, Smyth 2022). Regular practice test taking was also seen as important preparation (Baird 2019, Smyth 2022). Engaging teachers in assessment policy and reform was also seen as important for motivating teachers to deliver engaging teaching (Hepworth 2021).

In summary, student motivation varied depending on the system features, the purpose of the test, and the individual student.

Student stress and anxiety

This theme was related to the stress and anxiety experienced by students as a result of studying for and/or taking the evaluated assessment. Twelve studies were coded with this theme, and the results are presented in Appendix I.

Several studies reported students feeling pressure when preparing for or taking assessments that were 'high stakes' and especially assessments that would have consequences for students personally (Banks 2015, Brown 2022, McCoy 2018, Smyth 2022, and Zhao 2022). The main causes of pressure were the consequences to the student for their future (Banks 2015, Brown 2022, McCoy 2018, Smyth 2022, and Zhao 2022), the volume of content to cover (McCoy 2018) and competing demands (Smyth 2022). Of those studies reporting concerns, teachers and students noted that assessments, especially 'high stakes' assessment, was having ill-effects on student wellbeing, personal freedom and happiness (Hargreaves 2020, Smyth 2022). Jurisdictions where high-stakes assessment was not commonplace, reported lower levels of stress (Nieminen 2022). Stress and test anxiety was reported to be higher for female students (Banks 2015, Smyth 2022).

While some of this pressure was noted to be driven by students themselves (Banks 2015), teachers were noted to sometimes exacerbate these internal pressures (Banks 2015, Brown 2022, Smyth 2022). Although stress and test anxiety were considered negatively by some studies (Baird 2019, Brown 2022, Hargreaves 2020, McCoy 2018, Smyth 2022), some studies reported no association between test anxiety and attainment (Jerrim 2022) or had the opinion that "a little bit of stress is a good thing" (Lessof 2018, P.87).

Summary of findings about the study outcomes

- Evidence primarily from England and Ireland suggests that assessment systems and policy have a relationship with students' attainment, but in complex ways.
 - Facilitators of attainment include student positive perception of examination, the support of teachers, student choice or enthusiasm for the subject, doing separate science subjects rather than combined, doing homework, getting enough sleep, and school attendance.
 - Barriers to attainment include social economic status, strict threshold grading, and perceived poor linkage between the content of the taught subjects and the summative assessment.
- Student motivation varied depending on the system features, the purpose of the test, and the individual student. Teachers are likely to be important facilitators of student motivation.
- Stress and anxiety were commonly reported for high-stakes assessments, although a few studies did not find this negative association. The main causes of pressure were the consequences to the student for their future, the volume of content to cover, and competing demands. Teachers were noted to sometimes exacerbate these internal pressures.

Debates in the literature on summative assessment

Informing the context of the review work reported here were several debates around the benefits and challenges of different approaches to secondary education assessment. In the included papers for the review, we identified many debates which permeated the background context and discussions that are summarised below.

Potential harms to students

This was by far the most mentioned debate within the included papers. A lot of the ‘harms’ discussed in the included papers related to the perceived hardness of the assessment, stress / pressure / anxiety that those assessments caused students, and the resulting impact of high-stakes exams on young people’s emotional well-being (Baird 2019, Banks 2015, Brown 2022, Hargreaves 2020, Jerrim 2022, Putwain 2016). Some studies also raised concerns about the over-emphasis on students taking academic based forms of assessment, and the lack of choice students or lack of specialism with subject choice (Cadwallader 2013; Ofqual 2012). One study commented that harm comes from ‘the idea of rewarding those who perform best’ (McCoy 2018). An exception to this was the Finnish assessment system which was reported to “not include a strong tendency towards test anxiety” (Nieminen 2022, P.7). A perhaps unintended harm discussed is the affect assessments have on teaching, with ‘teaching to the test’ becoming a known phenomenon (Smyth 2012).

Implications for teaching practice

This debate centres around possible implications of assessments (current and future) for teachers and their practice. Baird (2019) comments that “the introduction of linear examinations, somewhat counterintuitively, appears to have led to more school assessments, particularly more mock examinations” (P.138), with the consequence of increased workloads for teachers. Some studies discuss the implications of the assessments on the curriculum. For example, Elwood (2017) discusses a narrowing of the curriculum to meet the demands of the assessment. Hepworth (2021) discusses Improving teacher quality as a renewed policy goal related to assessment. Machin (2018) argues that exam thresholds have also become increasingly important for incentivising teachers, especially in schools where pay is linked to school performance and rankings, and concerns are raised that this might lead to manipulation of marks by teachers. Putwain (2016) comments that teachers have an important role to play in helping prepare students with approaches for coping with the pressure of forthcoming examinations. Teacher support for assessments is especially key in high stakes assessment environments (Smyth 2012). Related to this point, Tong (2015) argues that teachers need to be skilled in providing feedback and enabling the students to make use of it to enhance their learning.

Implications for future assessment

Key themes here included the perception of the assessments by people external to the education system. For example, Baird (2019) found that teachers considered linear assessments to be ‘harder’ in terms of ‘rigour’ and credibility with the public and potential employers. Some of the discussion was around the concept of fairness, with Baird (2019) commenting that “some teachers felt that the ‘all or nothing culture’, created by the ‘high-stakes final, terminal exams’ would have a disproportionate impact on Pupil Premium students who wouldn’t have the best attendance, supportive family structures, or the cultural capital that would be needed” (P.121).

Hepworth (2021) cites the importance of designing future assessment in collaboration with key stakeholders, and “placing trust in schools to select the appropriate supports and interventions that make the most sense for their pupils” (P.63). Putwain (2016) argues that

supporting student with the development of 'academic buoyancy' could be an effective way of ameliorating the influence of worry or stress on attainment. Tong (2015) argues for the importance of including cultural contexts and embedding this into future assessment.

Reliability and validity of assessments

This debate was around the reliability and validity of current assessment types. Baird (2016) argues that "too much transparency can undermine learner autonomy and call into question the validity of inferences made from the assessments" (p. 5). Ofqual (2012) discusses the assessment format and discusses the use of multiple-choice questions "to test large amounts of content and assess skills such as the ability to use the English language" but are not commonly used for A level assessments, and questions if they should be (p. 234).

Predictability of content within assessments

Baird (2016) discusses how assessments can be criticised for being repetitive and predictable, which was argued to be a good tool for encouraging learning. Elwood (2017) comments on the mixed effect of predictability of content: "predictability is a dilemma because assessment transparency has positive effects, but if an assessment is overly predictable, it can have negative consequences for learning" (P.3). Elwood raises concerns about "teaching-to-the-test" and too much predictability leading to a "failure to develop a broad and deep understanding of subjects" among students (P.4).

High stakes assessment

Machin (2018) discusses the serious implications for students' future learning if they do not reach the required threshold grades. For example, Machin (ibid) argues that particular grades are needed at GCSE level to progress to A level, with selection criteria directly linked to performance. Smyth (2012) questions the effect of high stakes or national tests on teaching, learning and student outcomes in the Republic of Ireland. Zhao (2022) discusses how students will adjust their performance according to what is at stake with the assessment and prioritise individual consequence over national interest.

Relevance of disadvantage

Disadvantages may be in the process or outcomes of assessments perceived by students, and experiences by certain student groups, for example, those from low-income backgrounds, or between gender groups. Four studies were coded as addressing this theme.

Brown (2022) discusses the "unfairness" some students find with assessments. Lessof (2018) discusses the inequity with performance that is related to certain student groups undertaking assessments at KS4 and the longstanding effects this can have on social division. Machin (2018) discusses how students with the same ability can have different trajectories depending on the outcome of assessments, and the link of this to high stakes assessment. Debates on this topic highlight the importance of fairness and equality in determining the life trajectories of young people.

Student voice in policy reform

Two studies emphasised the importance of student voice in policy development or reform. Elwood (2013) discusses that students are often not consulted about the types of assessment to develop and yet when asked directly, they had clear views on it. For example, "students expressed concern with regard to how their achievements were being (de)valued, especially in relation to the currency of different qualifications" (p. 12). Elwood argues that student views around models and design of qualifications "seem to be at odds with what is being promoted by current government agendas for rigorous and robust qualifications" (p.

12). For these reasons Elwood supports students having a voice and a platform for raising concerns.

Breadth versus depth

Debates under this heading covered the types and appropriate levels of content that should be covered. For example, Ofqual (2012) discusses whether a high mathematical content for chemistry A level is good for students, or a broader approach might better prepare students better for studying in a higher education.

Summary of the debates identified in the literature

Key points of contention in the field of summative assessment in general lower and upper secondary education (ages 14-18/19), largely drawn from English and Irish evidence sources, were:

- Potential harms to the student, particularly in relation to stress and anxiety
- Implications for teaching practice, including workload, providing appropriate feedback, supporting students, narrowing the curriculum, and teaching to the test.
- Implications for future assessment, including design of future assessment with key stakeholders, ensuring fairness, and appreciating and embedding cultural context in assessment design.
- Reliability and validity of assessments, including the appropriateness of transparency and assessment format in assessment types.
- Predictability of content within assessments, which can have benefits (such as transparency for the student) and harms (such as narrowly focused learning and teaching to the test).
- High stakes assessments, which can have serious consequences for students' futures, and were often (but not always) associated with stress and anxiety.
- Relevance of disadvantage, which highlight the importance of fairness and equality in determining the life trajectories of young people.
- Student voice in policy reform, which is seen as crucial for developing assessments that are perceived as fair and valued.
- Breadth versus depth of curriculum content covered was emphasised as important in some studies, but it was not clear whether broader or more narrowly focused assessment systems are more desirable.

Conceptual framework of summative assessment

To help make sense of the diverse themes identified in the thematic synthesis, we constructed a conceptual framework that organised (i.e., configured) the themes. This is presented in Figure 1.

In the top left corner, we can see three red boxes that were stated purposes of the assessment at the school- or system-level. These were: to assess school level performance to inform parents' or carers' decisions about school applications, to incentivise teachers and/or school managers, and to assess school- or system-level performance for quality assurance or accountability reasons. (Note that, in the synthesis above, these three themes were merged into one higher theme around quality assurance and accountability). Also, within this larger box is it a purple box indicating a purpose to enable international comparisons, such as PISA. This is at a higher level of the system than the three aforementioned purposes, but it also has an accountability element and—importantly—they

all have implications for future assessment and future teacher practise (as emphasised in the adjacent grey boxes).

Moving across the figure to the top right corner, we see two themes in green boxes. The first relates to the purpose to ensure fairness or equality in assessment; the second is one of the debates that was raised throughout the papers around the relevance of disadvantage. The fact that this purpose was not always met was one of the reasons that it often came under criticism. This is partly linked to the sandy yellow box, which related to technical issues, including standardisation. One of the main reasons for standardisation and other technical aspects (e.g., establishing reliability and validity) was to ensure fairness and equality, but these processes were also sometimes noted to reduce inclusivity and exacerbate disadvantage.

The technical issues including standardisation also overlaps with some of the more specific themes around the form and content of the assessments in the bottom right corner of the figure. In the lighter blue boxes, we saw the issues of breadth versus depth of subject matter coverage and the predictability of content as being important themes for student motivation and whether students had learned the skills needed for their future life. Issues such as teaching to the test and narrowing of the curriculum to maximise grades cropped up here. These were seen as undermining one of the purposes of assessment, that being to assess whether learning objectives had been achieved and/or to assess student progress (the darker blue box).

To the bottom left corner, we see a cluster of orange boxes that relate to the purposes of the assessment regarding progression beyond the students' current schooling. Three of these related to moving on to other forms of education, those being higher levels of secondary school education, tertiary education, and vocational or technical education. The fourth orange box was for the purpose of allowing employers to differentiate between candidates. All four of these determine whether an assessment system is considered to be high or low stakes. The greater the emphasis on progression beyond current schooling, the higher the stakes.

At the bottom of the figure, we have two separate boxes: A lilac box relating to debates about the potential harms to students, and a pale turquoise box relating to calls for the need for student voice and policy reform. These were seen to underpin much of the rest of the framework. Any aspect of the assessment has the potential for to harm to students, while student voice in policy reform may mediate this impact.

In the middle of the figure is a darker-bordered box, containing our three main student outcomes: attainment, motivations to learn, and stress or anxiety. These were put at the centre of the framework because of the way in which the research studies generally took a student-centred approach. Overall, the research papers saw the other themes and debates raised here as potentially impacting on these three (and occasionally other) student outcomes.

The relations between all of these themes are complex, which is why we have not positioned lines of direction or indicators of magnitude of strength between them. Future research could look at the direction and strength of some of these associations. Also, as we will see in the comparative synthesis, this conceptual framework is developed very much with the research evidence perspective in mind. A broader systems perspective would look at impacts for different actors in this system, such as teachers, head teachers, assistant assessment system managers, and policy makers.

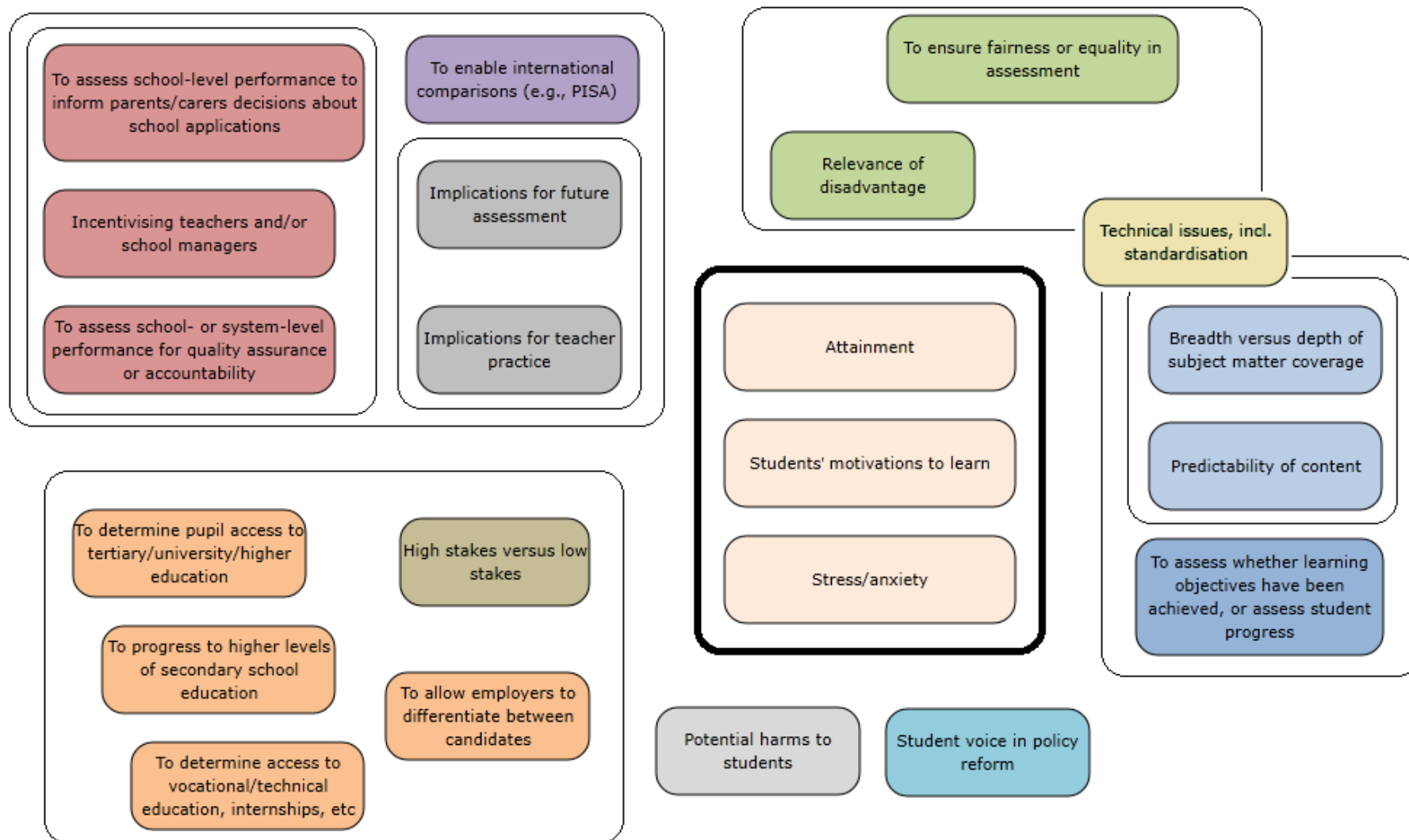


Figure 1. Conceptual framework of summative assessment derived from the thematic synthesis of research evidence.

Results of the comparative synthesis

After reviewing the ten individual jurisdiction reports, five major themes were identified across the policy summaries:

- **the balance of formative versus summative assessment** within jurisdictions, which was often linked to concerns about high-stakes testing. For instance, in Finland, the key consideration for education policymakers and educators is the balance between two assessments as they aim to promote effective learning and ensure fair evaluation of students' knowledge and skills.
- **the amount of student choice/flexibility within a system**, which included different pathways or types of qualification or certification that students could attain. Understanding that students have varied interests, strengths, and career aspirations, jurisdictions such as Ontario have sought to offer different pathways or types of qualifications/certifications that students can attain. This approach acknowledges that not all students follow a single, standardised path and encourages personalised learning experiences that cater to individual needs and goals.
- **the timing and extent of assessment policy reform**. Policymakers have recognised the requirement to adapt assessment practices to meet the growing demands of the 21st-century education landscape. Reforms have intended to align assessment methods with the anticipated learning outcomes and pedagogical approaches, fostering a holistic and comprehensive assessment of student performance. The timing and extent of these reforms have varied across jurisdictions in the European Union and overseas, reflecting local contexts, educational ideologies, and the readiness to embrace change.
- **an emphasis on fairness and inclusivity in recent reforms**. In all jurisdictions, there is a rising recognition of the prominence of delivering equal opportunities for all students, regardless of their background or circumstances. Efforts have been made to promote equity in assessment processes, ensuring that students from different backgrounds have attained a level of achievement for themselves.
- a strong emphasis on the labour market and international competitiveness, with varying levels of motivation towards these external influences.

These themes were overlaid with the themes identified in the research evidence (see Table 7). From this comparative synthesis, we conclude that the research evidence base may be under-representing some of the key themes that may be of interest to policymakers. In contrast, research papers tended to focus more on the experiences or outcomes for students or teachers; for example, policy reports did not have the same emphasis on concerns about student stress and anxiety as in the research papers. The policy reports were also unable to capture nuances about fairness and equality as expressed by research participants in the reviewed research studies. These differences may be because of the study identification methods that we followed in the review, or it may be that the researchers tend to focus on individual level outcomes for students or teachers, whereas policymakers tend to focus on national or international-level concerns.

It is, however, clear that educational assessment policy is part of a broader system of external influences, including the jurisdiction's education system, local and international labour markets, and international competition, as well as internal factors, including students, teachers, headteachers, schools, and the assessment/certification agency. Each of these influences and factors were touched on to differing extents in the research and policy documents, but as noted above, the research tended to look at policies impacts from the bottom-up, whereas policy documents take a more top-down perspective.

Table 7. Observations of the coverage of policy themes in the research evidence in this review

Themes across the policy reports	Amount and type of coverage in the research in this review	
The balance of formative versus summative assessment	Low	This only received light coverage in the research studies. Most of the assessments covered were high-stakes, terminal assessments.
The amount of choice/flexibility within a system, which included different pathways or types of qualification/certification that students could attain	Low	This received moderate attention in the background sections of the papers but was not the focus of evaluations identified here.
The timing and extent of assessment policy reform	Moderate	Reactions to policy reform was reported, including both student and teacher perceptions about the introduction of features including school-based assessments, linear assessments, and higher-level offers of some subjects.
An emphasis on fairness and inclusivity in recent reforms	Good	This was covered in several studies, with mixed findings as to whether fairness and equity were achieved. Student voice and stakeholder engagement were seen as important to achieve this in the future.
A strong emphasis on the labour market and international competitiveness.	Low	This was indirectly addressed in our observation that the stated purposes of the assessment were not always the same as the stated uses of the assessment, with <i>use</i> by the labour market in recruitment selection processes quite common even though it was not often an intended <i>purpose</i> acknowledged by the studies. A growing realisation that labour market needs are important in policy design may lead to different types of assessment, possibly with a greater focus on skills. International comparisons did arise in the research review, but additional relevant studies may have been identified if the search strategy had been developed with this as its focus.

Note. The evidence base was largely comprised of evidence from England and Ireland, possibly because of the limitation on English language. Different coverage may be evident in other countries/languages.

Results of the comparison tables

The following tables (see Tables 8-12) provide a descriptive analysis 'at a glance' of the key policy differences and similarities between the jurisdictions.

RQ1: What is the current summative assessment system(s)?

The results for this research question are presented in Table 8 towards the end of this section.

Stages of formal assessment

Eight of the ten jurisdictions clearly divide upper secondary into a lower stage and an upper stage that each have their own qualification (the exceptions are Ontario and Japan, which are discussed further below). Across jurisdictions, the lower stage typically ends at around 15 or 16 years of age, with the exception of Ontario, which runs through to 18 years. The end of the lower stage typically coincides with the end of compulsory education, except for England, Finland, and Poland, where compulsory education or training continues until 18 years of age.

The two exceptions to the two-stage system are Ontario and Japan. In Ontario, the lower stage is not distinguished; rather, it forms a part of what is referred to as a 'continuous educational system' for ages 14/15 through to 18. In Japan, a lower stage is marked by a National Assessment of Academic Ability at around age 15 years, but there is no associated qualification, while at the upper stage, there is currently no national summative assessment system.

At the upper stage, the same observations at the lower stage generally hold, with the exceptions of Poland and Hong Kong. In Poland, students can exit with different qualifications depending on the type of assessment taken. In Hong Kong, the system is designed to provide various pathways to higher education and the workplace.

Vocational pathways

All jurisdictions offer a **vocational pathway** or route, typically at the post-compulsory upper stage, apart from Japan. In Japan, vocational courses can be selected, but it is not clear that a distinct qualification is awarded upon completion.

Subject choice

The jurisdictions varied widely in terms of the **number of core (compulsory) subjects and elective (optional) subjects** that could be taken. There was a minimum of three core subjects at the lower stage, with some jurisdictions making all lower stage subjects compulsory. At the upper stage, the number of compulsory subjects varied from zero to up to 18. Finland stood out by its inclusion of at least one multidisciplinary learning module that must be determined locally at the lower stage, while Ontario's upper stage requires additional mandatory 40 hours of community service.

The assessments

Some countries have limited **assessment approaches** (e.g., England's system is dominated by one assessment at each stage). Others have a wide range of assessments at

each stage and include much more of a mix between formative and summative assessments.

Most of the systems had a mix of **assessment forms**, including examination, project work, and school based or teacher assessments. Japan and Hong Kong have only examination as the form of summative assessment. At the upper stage in all jurisdictions, there appeared to be some form of final examination as the primary assessment method, although all offered a mix of other assessment forms. Because of its credit system and lack of a national summative assessment system at the upper stage, Japan has a more flexible approach to assessment at the upper stage and will vary depending on the credits selected by the student.

Most systems had either fully or partially **standardised** assessments. The exceptions were Poland, where examinations are not based on a common set of questions, and New Zealand, in which there is no single set of exams that all students take at the same time. Instead, New Zealand's system is flexible and allows for different pathways depending on the interests or perceived strengths of the student. For Ontario, there was conflicting evidence about whether it was standardised or not.

In terms of the **mode of assessment**, most systems reported a mix of paper and electronic assessments, although generally more paper based. Estonia, Hong Kong, and Japan are currently exclusively paper-based but are considering a shift to electronic assessment. In Finland, the matriculation exam is fully electronic.

RQ2: What are the stated purposes of each summative assessment?

The results for this research question are presented in Table 9 towards the end of this section. Of the ten jurisdictions, there was evidence that:

- Seven used assessment results for teacher training and/or to improve the curriculum,
- Eight jurisdictions used results for regional and system accountability,
- Six used results to hold schools accountable, in order to improve schools or regions, and/or to inform allocation of resources,
- Seven jurisdictions clearly **publish results at the institution and regional level**. In Estonia, secondary school assessments are not published in league tables or official statistics. In Hong Kong, TSA results are only published at a territory-wide level, and the results of HKDSE are released at territory-wide level but not by individual school. In Japan, national averages and mean subject scores for each of the 7 regions of Japan and ranks the 47 prefectures accordingly, but institution-level results are not necessarily published since the decision as to whether to publish is delegated to education policymakers in each of the 1718 municipalities of Japan. Each municipality can decide whether to publish individual schools' mean scores. In the cases of Hong Kong and Japan, the purpose of summative assessment is more about providing information to improve school performance and identifying system-level changes that are needed, rather than focusing on individual student outcomes (e.g., progression to further education).

RQ3: What is the process through which national-level assessment policy is made?

The results for this research question are presented in Table 10 towards the end of this section.

In all cases, the national education department makes and usually **specifies the assessment policy**. Typically, a different organisation has a specific remit to **regulate** how the policy is being implemented regionally and at school level. In most cases, this an independent organisation from government.

Seven jurisdictions **explicitly encourage local interpretation and/or implementation** of the nationally set assessment policies. For example, Finnish teachers have independence on how to assess the core curriculum content, while the Republic of Ireland emphasises the 'partnership approach'. In the Netherlands, schools have autonomy to tailor their educational assessments to meet the needs and interests of their students and communities. Explicit encouragement of local implementation was not evident in England, while it was only partially evident in Poland and unclear in Japan.

RQ4: Through what mechanisms is the current summative assessment system evaluated?

The results for this research question are presented in Table 11 towards the end of this section.

Evaluation mechanisms are varied and include research, consultation with stakeholders (e.g. Ontario's framework includes stakeholder involvement), monitoring, and a mixture of other internal and external processes. The main purposes seem to be: to check the quality of provision and adherence to the policy, to check for outcomes and effectiveness of the assessment system, and to compare educational outcomes for students, schools, and regions.

In eight of the ten jurisdictions, a government body is **responsible for evaluation** of the summative assessment system. A notable exception is Estonia, where each school and its teachers are responsible for evaluating the quality of the assessments that they use. In the Republic of Ireland, there are variety of stakeholders involved in evaluation in addition to a government body. In Japan, there are two associations that have published test standards and codes of practice; however, they are not directly involved in decision-making on national summative assessment policy.

RQ5: What contextual factors are reported to have driven national-level assessment policy?

The results for this research question are presented in Table 12 towards the end of this section.

Common **internal drivers of policy** (i.e., identified by actors within the assessment/curriculum process) that were identified include concerns about educational

outcomes, and addressing inequalities, fairness, and equity. Other themes that emerged included: meeting the needs of a diverse range of students, aligning assessments with curriculum standards, modernising the skills and competencies being taught and assessed, and meeting the requirements of a global and local employment market.

Concerns specific to some countries include:

- transparency of mechanisms and processes (Poland),
- inclusivity and fairness (Estonia, Poland, the Republic of Ireland, the Netherlands, Ontario and New Zealand),
- the intense pressure that students are under (Hong Kong and Japan),
- the possibilities of cheating or 'gaming the system' (England),
- using more formative assessment or a mix of formative and summative, and personalising the student journey (Estonia, the Netherlands, the Republic of Ireland and New Zealand).

External drivers of policy (i.e., influencing factors outside the assessment or/curriculum process) that were identified include:

- Competing in a global economy (England, Estonia, Poland, the Netherlands, Japan),
- Historical influences from Chinese and British education systems in Hong Kong
- Confucian heritage influence in Japan,
- Concerns about inequity introduced by differential access to expensive tutoring schools (*juku*) in Japan,
- Aligning with EU requirements (Poland)
- The COVID-19 pandemic necessitated a balance between assessment, expectations and student flexibility and support (England) and integrating technology into classrooms (Japan) or assessment (Finland) sooner than planned.
- Responding to PISA results and other international comparisons (Republic of Ireland and Hong Kong).

Overarching observations

From the table and the textual commentary, what stands out is that every jurisdiction has its own ways of doing things, although there are some parallels. The parallels are particularly evident between the EU countries, in terms of aspirations to deliver a system that values formative feedback and recognises student individuality.

Most systems offer a two-stage approach (except Ontario and Japan), and compulsory education ends around 15/16 (except Ontario and Poland, where it is 18). Japan stands out most of all with a fully credit-based system. Vocational pathways do not differentiate between jurisdictions as anticipated; all jurisdictions offer them in at least some form, and nine of the ten offer a formal VET programme.

It is difficult to determine whether a given jurisdiction has high, medium, or low stakes from the perspective of the student. Most systems offered a mixture of different assessment forms, typically culminating in a final examination. The emphasis placed on the final exam may differ; for example, England places heavy emphasis on the final exams, whereas Finland, the Republic of Ireland, Ontario, and New Zealand seem to give more relative weight to other assessment types. This arguably reduces the personal stakes for the students. Hong Kong and Japan are also arguably lower stakes systems because, despite being exam-only at the lower stage, the implications of performance are at the school- and system-levels rather than for the individual students; however, this interpretation would be at odds with other known pressures in those systems around cultural expectations of student performance.

Evaluation of the systems is formalised in all jurisdictions, yet seven of the jurisdictions encourage local implementation of the policies. This latter point allows for some flexibility in meeting local needs.

A host of both internal and external drivers have been proposed as influencing summative assessment policy across jurisdictions. Broadly, concerns about wellbeing, inclusivity and fairness, modernisation, and competing in a global economy stood out as recurring themes. Individual jurisdictions, however, had specific influences that drove specific policies that may not transfer to other contexts.

Finally, we observed differences across jurisdictions regarding whether the system looked at the student more holistically, by considering broader aspects of their personal development or student experience or took a narrow lens to what education is about, by focusing more on the academic aspects of learning. Examples of attempts to take a broader perspective include Hong Kong, which conducts an Attitudinal and Learning Experience Questionnaire alongside the academic assessments to gauge student experiences, and Ontario, which requires community service as part of the assessable curriculum. Some systems had more of an emphasis on diversity, inclusion, and meeting the needs of the children (e.g., through broader skill development), rather than system level performance or narrow academic foci. Across most systems, however, there appeared to be a growing move towards a system that reduced stress and increased equity for the students. How these ambitions are enacted in future policy reforms remains to be seen.

Table 8. Jurisdiction-level summary information about secondary school assessment policy and practice for research question 1 (correct as of November 2023)

RQ1: What is the current summative assessment system(s)?

	England	Finland	Estonia	Poland	Republic of Ireland	The Netherlands	Hong Kong	Ontario	Japan	New Zealand
Stages of formal assessment	Lower stage: GCSE (14-16y)	Lower stage: Basic (14-16y)	Lower stage: Basic School State Examination (up to age 16)	Lower stage: (aged 13- 16 years) none ¹	Lower stage: Junior Cycle Examination (12-15 years)	Lower stage: Summative assessments are conducted during age 12-16, culminating in the VMBO (preparatory secondary vocational education) ²	Lower stage: TSA, (14-15 years) ³	Lower stage: (ages 14/15-16) none ⁴	Lower stage: National Assessment of Academic Ability (NAAA) (15 years)	Lower stage: NCEA ⁵ at final year of compulsory schooling (Year 11, age 15)
	Upper stage: AS and A-stage (17-18y)	Upper stage: General Upper Secondary Education (17-18y)	Upper stage: Upper secondary school leaving certificate (16-19y)	Upper stage: 'Matura' exam (age 18y) ⁶	Upper stage: Senior Cycle of Examination (15 to 18 or 19 years) End is the Irish Leaving Certificate examination	Upper stage: The Secondary School Leaving Examination (Examen Voortgezet Onderwijs) (17-18y)	Upper stage: Hong Kong Diploma of Secondary Education (HKDSE) (17-18 years)	Upper level: grades 9 to 12 ⁷	Upper stage: none ⁸	Upper stage: NCEA Level 2 takes place in Year 12 (age 16) and Level 3 in Year 13 (age 17) ⁹

	England	Finland	Estonia	Poland	Republic of Ireland	The Netherlands	Hong Kong	Ontario	Japan	New Zealand
Is there a technical or vocational secondary-level pathway available?	Yes ¹⁰	Yes ¹¹	Yes ¹²	Yes ¹³	Yes ¹⁴	Yes ¹⁵	Yes ¹⁶	Yes ¹⁷	Unclear ¹⁸	Yes ¹⁹
Last age of compulsory education	16y for compulsory full-time education; 18y for compulsory education or training ²⁰	18y ²¹	16y	18y	16y	16y ²²	15y	18y	15y	16y
Are subjects compulsory, elective, or a mix?	Lower stage: GCSE (9 usually taken, 3 core subjects – maths, English, and science - plus 6 electives), BTEC (cover a range of	Lower stage: up to 20 compulsory subjects, plus at least one multidisciplinary learning module	Lower stage: up to 7 core, and up to 12 electives	Lower stage: all subjects are compulsory (up to 18 subjects)	Lower stage: 10 subjects, 7 core, and 2 optional	Lower stage: Mixed ²³	Lower stage: three compulsory subjects (Chinese language, English language, and maths) are examined, although	Lower stage: Subjects in lower secondary school in Ontario are compulsory and follow a set curriculum ²⁴	Lower stage: 4 core subjects (maths, Japanese, science, and English)	Lower stage: Level 1 NCEA subjects are compulsory (English, Maths and Science)

	England	Finland	Estonia	Poland	Republic of Ireland	The Netherlands	Hong Kong	Ontario	Japan	New Zealand
	subjects, some core, others optional)	determined locally					more maybe studied			
	Upper stage: 3-4 subjects (pupil choice), BTEC (cover a range of subjects)	Upper stage: 4 compulsory and 2-4 elective subject areas	Upper stage: up to 7 core, and up to 12 electives	Upper stage: the basic level is compulsory (up to 18 subjects) and the advanced level are optional for students (up to 10 subjects)	Upper stage: typically, students take six or seven subjects	Upper stage: Same as lower	Upper stage: 4 core and 2-3 electives	Upper stage: A mix of compulsory and elective subjects ²⁵	Upper stage: Unspecified – depends on the credits selected	Upper stage: Level 2 and 3 NCEA is not compulsory ²⁶
What forms do assessments take (e.g., teacher assessment, examination, coursework) and when	Lower stage: Mix: Largely examinations, with elements of project work, and practical demonstration	Lower stage: Mix: Examination, project work, practical demonstration	Lower stage: Mix: Examination, project work, practical demonstrations Includes	Lower stage: Mix, Examination, project work, presentations, portfolios	Lower stage: Mix, Examination, project work, and teacher assessments	Lower stage: Mix, Examination, project work, and teacher assessments	Lower stage: TSA is exam-based (Basic Competency Assessment) plus a	Lower stage: A mix of tools and approaches, including tests, examinations, projects,	Lower stage: Examination	Lower stage: Mixed Internal assessments take various forms, from coursework,

	England	Finland	Estonia	Poland	Republic of Ireland	The Netherlands	Hong Kong	Ontario	Japan	New Zealand
does it take place?	tions in some subjects Main assessment (exam) is terminal	s Can vary by local context and student	self- and peer-assessment	observati ons			questionnaire (Attitudinal and Learning Experience Questionnaire)	portfolios, and observati ons		projects, presentati ons, or other forms External are exams
	Upper stage: Same as lower	Upper stage: Same as lower, but end exam	Upper stage: Same as lower, but end exam	Upper stage: Same as lower	Upper stage: same as lower, but end exam	Upper stage: Same as lower	Upper stage: Mix: examination, school-based assessments	Upper stage: Same as lower	Upper stage: Unspecifie d – depends on the credits selected	Upper stage: Same as lower
What is the balance between normative and criteria based	Lower stage: Adjusted within a cohort and some normalising, but also some criteria that need to be met	Lower stage: Appears to be primarily criteria-based	Lower stage: A grading system is used to classify the outputs of summative assessment	Lower stage: criteria based It uses scoring criteria	Lower stage: a combination of normative and criteria based	Lower stage: VWO is overall criterion-referenced with normative elements	Lower stage: criteria-based	Lower stage: rubrics outline the criteria for assessing different aspects of writing, such as organisation,	Lower stage: Primarily criteria-based	Lower stage: more criteria based

	England	Finland	Estonia	Poland	Republic of Ireland	The Netherlands	Hong Kong	Ontario	Japan	New Zealand
								content, and language conventions		
	Upper stage: Same as lower	Upper stage: A grading system is used to classify the outputs of summative assessment	Upper stage: Same as lower	Upper stage: mixed Depending on the type and purpose of the assessment, it can be either criteria-based or normative Internal assessment is criteria based, external ('Matura') is normative	Upper stage: mainly criteria based but some elements of normative (Transition Year programme is offered by some schools as an optional year between the Junior Cycle and the Senior Cycle)	Upper stage: similar to the lower level Generally speaking more normative	Upper stage: Same as lower	Upper stage: Same as lower	Upper stage: Unspecified because there is no national summative assessment system at this level	Upper stage: Same as lower
Mode of assessment	Lower stage:	Lower stage:	Lower stage:	Lower stage:	Lower stage: A	Lower level: both paper	Lower stage:	Lower stage: A	Lower stage:	Lower stage:

	England	Finland	Estonia	Poland	Republic of Ireland	The Netherlands	Hong Kong	Ontario	Japan	New Zealand
	Mixed, but mostly paper-based	Mixed, but matriculation exam is fully electronic	Currently paper based but considering transition to electronic	some are paper-based and some are electronic	combination of paper and electronic	based and electronic, depending on the type and form of the assessment	paper-based	combination of paper and electronic	paper-based	assessments can be on paper or electronic Move to electronic being encouraged
	Upper stage: Same as lower/	Upper stage: Same as lower	Upper stage: Same as lower	Upper stage: Same as lower	Upper stage: mainly paper	Upper stage: Same as lower	Upper stage: Same as lower	Upper stage: Same as lower	Upper stage: Unspecified because there is no national summative assessment system at this level	Upper stage: Same as lower
It is standardised? If so, at what stages, and is it at the local/regional or national level?	Lower stage: Yes for GCSE; no for BTECs	Lower stage: Yes, for the last year of school exam ²⁷	Lower stage: Yes	Lower stage: no, it is not based on a common set of questions	Lower stage: partially ²⁸	Lower stage: partially ²⁹	Lower stage: Yes	Lower stage: Unclear	Lower stage: Yes	Lower stage: No ³⁰

	England	Finland	Estonia	Poland	Republic of Ireland	The Netherlands	Hong Kong	Ontario	Japan	New Zealand
	Upper stage: Same as lower	Upper stage: Same as lower	Upper stage: Same as lower	Upper stage: mixed ³¹	Upper stage: Yes ³²	Upper stage: No ³³	Upper stage: Yes	Upper stage: Yes	Upper stage: Unspecified ³⁴	Upper stage: Same as lower
When was the current system introduced (based on the last major policy reform)?	Lower stage: GCSE: 2017, 1996 for BTECs but new reform expected for 2024	Lower stage: 2017	Lower stage: 2014	Lower stage: 1999 after major reform	Lower stage: 1990s	Lower stage: 1999 with reforms to the Mammoth Act	Lower stage: 2009	Lower stage: Unclear	Lower stage: 2013 (with science added in 2015 and English in 2019)	Lower stage: 2023
	Upper stage: AS and A-levels: 2017, same for BTECs	Upper stage: 2016	Upper stage: 2013	Upper stage: 2019 as part of a reform of education	Upper stage: same as lower	Upper stage: same as lower	Upper stage: 2012	Upper stage: same as lower	Upper stage: Unspecified ³⁵	Upper stage: Differs according to level of assessment ³⁶

Table 9. Jurisdiction-level summary information about secondary school assessment policy and practice for research question 2 (correct as of November 2023)

RQ2: What are the stated purposes of each summative assessment?

Stated purposes	England	Finland	Estonia	Poland	Republic of Ireland	The Netherlands	Hong Kong	Ontario	Japan	New Zealand
To improve teacher training and/or curriculum	No ³⁷	Yes	Yes ³⁸	Yes	Yes ³⁹ uses of lower secondary school assessments in Republic of Ireland	Yes ⁴⁰	Yes ⁴¹	Unclear	Yes ⁴²	Unclear
For regional and system accountability	Yes for both GCSEs and A levels ⁴³	Yes ⁴⁴	No ⁴⁵	Yes ⁴⁶	Yes ⁴⁷	No	Unclear	Yes ⁴⁸	Yes ⁴⁹	Yes

Stated purposes	England	Finland	Estonia	Poland	Republic of Ireland	The Netherlands	Hong Kong	Ontario	Japan	New Zealand
To improve schools/regions and aid allocation of resources	No ⁵⁰	No ⁵¹	Unclear	Yes	Unclear	Yes ⁵²	Yes ⁵³	Unclear	Yes ⁵⁴	Yes ⁵⁵
For tracking or sorting of students – lower secondary	Unclear ⁵⁶	Unclear	Yes	No ⁵⁷	Yes	Yes	Yes	Yes ⁵⁸	No ⁵⁹	Unclear ⁶⁰
For tracking or sorting of students – upper secondary	Unclear ⁶¹	Unclear	Yes ⁶²	Unclear	Yes	Yes	Yes ⁶³	Yes ⁶⁴	Unclear	Unclear
Are the aggregated results published, and if so, are they published at institution or regional level?	Yes	Yes	No ⁶⁵	Mixed ⁶⁶	Yes	Yes	Yes ⁶⁷	Yes	Yes ⁶⁸	Yes ⁶⁹

Table 10. Jurisdiction-level summary information about secondary school assessment policy and practice for research question 3 (correct as of November 2023)

RQ3: What is the process through which national-level assessment policy is made?

	England	Finland	Estonia	Poland	Republic of Ireland	The Netherlands	Hong Kong	Ontario	Japan	New Zealand
Who makes/specifies the policy at national level?	Department for Education	The National Board of Education (Opetushallitus) ⁷⁰	Central government ⁷¹	The Ministry for Education and Science ⁷²	Minister of Education, supported by the Department of Education	Ministry of Education, Culture and Science	Education Bureau (EDB)	Ontario Ministry of Education	Ministry of Education, Culture, Sports, Science and Technology (MEXT)	The Ministry of Education (MoE) ⁷³
Who writes and regulates the policy at national level?	Ofqual ⁷⁴	The Finnish National Agency for Education ⁷⁵	Central government ⁷⁶	The Central Examination Board (CEB) ⁷⁷	The National Council for Curriculum and Assessment (NCCA) ⁷⁸	Ministry of Education, Culture and Science ⁷⁹	Hong Kong Examinations and Assessment Authority (HKEAA) ⁸⁰	Ontario Ministry of Education ⁸¹	Ministry of Education, Culture, Sports, Science and Technology (MEXT)	The New Zealand Qualifications Authority (NZQA)
Is there explicit encouragement of local (e.g., regional, school, or pupil level) implement	No ⁸²	Yes	Yes	Some ⁸³	Yes ⁸⁴	Yes ⁸⁵	Yes ⁸⁶	Yes	Unclear	Yes

ation of the curriculum and/or assessment?										
--	--	--	--	--	--	--	--	--	--	--

Table 11. Jurisdiction-level summary information about secondary school assessment policy and practice for research question 4 (correct as of November 2023)

RQ4: Through what mechanisms is the current summative assessment system evaluated?

	England	Finland	Estonia	Poland	Republic of Ireland	The Netherlands	Hong Kong	Ontario	Japan	New Zealand
What are the evaluation mechanisms?	Research and analysis; consultation; monitoring and oversight; standard-setting; and evaluation and review	Various stakeholders ⁸⁷	Results are compared across different types of schools ⁸⁸	Evaluation and accreditation processes ⁸⁹	Individual school monitoring involving various internal stakeholder groups ⁹⁰	internal and external evaluation processes ⁹¹	A regular stakeholder review and consultation process	Accountability monitoring ⁹²	School and parent questionnaires ⁹³	Formal quality assurance processes ⁹⁴

Who is responsible for evaluation?	England Ofqual	Finland Finnish National Agency for Education (EDUFI)	Estonia Schools and teachers	Poland The Polish Accreditation Committee and the General Council ⁹⁵	Republic of Ireland Variety of Stakeholders ⁹⁶	The Netherlands The Inspectorate of Education	Hong Kong The EDB for the TSA; the HKEAA for the HKDSE	Ontario The Education Quality and Accountability Office (EQAO)	Japan Central Council for Education ⁹⁷	New Zealand NZQA
------------------------------------	-------------------	--	---------------------------------	--	--	--	---	---	--	---------------------

Table 12. Jurisdiction-level summary information about secondary school assessment policy and practice for research question 5 (correct as of November 2023)

RQ5: What contextual factors are reported to have driven national-level assessment policy?

	England	Finland	Estonia	Poland	Republic of Ireland	The Netherlands	Hong Kong	Ontario	Japan	New Zealand
Internal drivers of policy (identified by actors within the assessment/ curriculum process)	Qualification standards and technology ⁹⁸	Education al inequalities ⁹⁹	The need for core competence s ¹⁰⁰	Education al outcomes ¹⁰¹	Educational outcome s and justice ¹⁰²	Quality assurance and equity ¹⁰³	Pressure on students and to provide options for students ¹⁰⁴	Transpare ncy and fairness ¹⁰⁵	Education al outcomes ¹⁰⁶	Inequalities ¹⁰⁷
External drivers of policy (influencing factors outside the assessment/ curriculum process)	Competin g in a global economy ¹⁰⁸	Social justice culture ¹⁰⁹	Competing in a global economy ¹¹⁰	Competin g in a global economy ¹¹¹	PISA and other international standard s	Competing in a global economy	Global education al influences ¹¹²	None mentioned in the brief	Competin g in a global economy, cultural factors ¹¹³	None mentioned in the brief

Discussion

Summary of findings across the research

This section brings together the summaries from the different parts of the report in one place.

Description of the evidence on summative assessment in general lower and upper secondary education (ages 14-18/19)

- The review included 21 primary research studies from seven jurisdictions. Most of the studies were from England (n = 12). This was followed by the Republic of Ireland (n = 7); Finland, Hong Kong, and New Zealand (n = 2 each); and Ontario and The Netherlands (n = 1 each). There were no eligible studies identified for Estonia, Poland, and Japan.
- The most reported *purpose* of the assessment system was to assess whether learning objectives have been achieved (n = 9), then to enable progression through levels of education (n = 8).
- The most reported *use* of the assessment system was to enable progression through levels of education (n = 14), with the majority of those focused on progression to higher education (n = 12). The next most common use was to allow employers to differentiate between candidates (n = 11).
- Most of the studies explicitly stated that the assessment used examinations to assess the required standard (n = 15), sometimes in combination with other assessment types such as coursework.

Summary of findings about purposes of the assessment

- Assessment systems do achieve the purpose of allowing progression to the next stage of education, whether it be upper secondary, tertiary, or vocational education.
- Systems were generally found to capture individual student learning well.
- There was insufficient evidence to determine whether the assessment policies enable quality assurance or auditing of the quality of schools or the education system.
- PISA may be used for international comparisons of education systems in terms of student performance, but whether it adequately achieves this purpose is not answerable in this review due to the study identification methods used.
- The purpose of ensuring fairness and equality was not always met, with concerns raised about gender stereotyping, teacher bias, subject bias, and inequalities in the grading structures.
- Concerns were also raised about whether the assessment systems:
 - equipped the students with the learning they needed in the future, particularly practical skills.
 - overemphasised maximising grades at the expense of learning
 - risked limiting short- to medium-term opportunities if students failed to pass threshold grades.
 - become overly complex when the system combines formative and summative elements.
 - were adequately explained to the students.
 - prioritised certain core academic subjects (such as maths) over other 'marginal' subjects (such as art) in ways that could be unfair to students with skills or interests in the latter.

- Facilitators to improved learning included:
 - providing helpful, relevant feedback.
 - providing content specific to the subject being taught.
 - including student voice when designing or revising secondary level assessments.

Summary of findings about the study outcomes

- Assessment systems and policy appear to have a relationship with students' attainment, but in complex ways.
 - Facilitators of attainment include: student positive perception of examination, the support of teachers, student choice or enthusiasm for the subject, doing separate science subjects rather than combined, doing homework, getting enough sleep, and school attendance.
 - Barriers to attainment include: social economic status, strict threshold grading, and perceived poor linkage between the content of the taught subjects and the summative assessment.
- Student motivation varied depending on the system features, the purpose of the test, and the individual student Teachers are likely to be important facilitators of student motivation.
- Stress and anxiety were commonly reported for high-stakes assessments, although a few studies did not find this negative association The main causes of pressure were the consequences to the student for their future, the volume of content to cover, and competing demands Teachers were noted to sometimes exacerbate these internal pressures.

Summary of the debates identified in the literature

Key points of contention in the field of summative assessment in general lower and upper secondary education (ages 14-18/19) were:

- Potential harms to the student, particularly in relation to stress and anxiety.
- Implications for teaching practice, including workload, providing appropriate feedback, supporting students, narrowing the curriculum, and teaching to the test.
- Implications for future assessment, including design of future assessment with key stakeholders, ensuring fairness, and appreciating and embedding cultural context in assessment design.
- Reliability and validity of assessments, including the appropriateness of transparency and assessment format in assessment types.
- Predictability of content within assessments, which can have benefits (such as transparency for the student) and harms (such as narrowly focused learning and teaching to the test).
- High stakes assessments, which can have serious consequences for students' futures, and were often (but not always) associated with stress and anxiety.
- Relevance of disadvantage, which highlight the importance of fairness and equality in determining the life trajectories of young people.
- Student voice in policy reform, which is seen as crucial for developing assessments that are perceived as fair and valued.
- Breadth versus depth of curriculum content covered was emphasised as important in some studies, but it was not clear whether broader or more narrowly focused assessment systems are more desirable.

Summary of findings from the comparative synthesis

- The research sources tend to focus on individual level outcomes for students or teachers (bottom-up), whereas policy sources tend to focus on national or international-level concerns (top-down).
- A key area of overlap was around the desire for fairness and inclusivity for students, which appears to not always be met; involvement of student voice in policy reform was highlighted as an important strategy for addressing this.

Transferability to the English context

Each jurisdiction that we examined in this project differed in so many ways, making it challenging to unpick which factors might be the most important for understanding transferability to England. We have, however, some clues from the policy sources about how some of the systems differ to the English context.

Perhaps the closest jurisdiction—geographically, culturally, and systemically—to the English context is the Republic of Ireland. It is also the jurisdiction where most of the research evidence came from after England. This makes it particularly important to consider transferability of the Irish assessment system to England.

In Ireland, there is a focus on the assessment system's continual evaluation and monitoring in order to maintain its effectiveness and adaptability to changing educational settings. These contextual factors highlight the importance of connecting assessment practises with curriculum goals, focusing on important skills and abilities (such as critical thinking, problem-solving, communication, collaboration, creativity, and digital literacy), and exploring possibilities for continuous assessment approaches when considering transferability to the English education system. According to the OECD (2020)⁸ report, understanding the experiences and lessons learned from Ireland's assessment reforms might provide useful insights for England's own assessment policy and reforms, allowing for a more student-centred and holistic approach to assessment.

Finland is also an important system to consider, given the lower reported stress of students – purportedly due to the more balanced mix of formative and summative assessments. The Finnish model of national-level assessment policy and reforms may not be entirely transferable due to differences in their education systems and societal contexts. However, some of the potential lessons highlighted in the sources that could be relevant to assessment policy and reforms in England include:

- *teacher autonomy and trust*: Finland's education system is based on a high level of trust in teachers and a belief that they are the best reviewers of how to educate their students. This has led to a system where teachers are given a lot of autonomy and professional freedom, which could be relevant to assessment policies that aim to empower teachers to make decisions about their own assessments.
- *formative assessment*: Finland places a strong emphasis on formative assessment, or assessment that is used to inform and improve learning as it happens, rather than simply evaluating it at the end. This could be relevant to assessment policies that aim to prioritise learning over grading.
- *equity*: Finland has a strong commitment to equity and reducing inequality in education outcomes. This could be relevant to assessment policies that aim to

⁸ OECD. (2020). Education in Ireland. Retrieved from <https://www.oecd-ilibrary.org/sites/7b915e09-en/index.html?itemId=/content/component/7b915e09-en>

address issues of educational inequality, such as by providing support and resources to schools in disadvantaged areas.

As another near neighbour to England, the Netherlands may also have important lessons for transferability. The reforms and standards for assessment and its monitoring may be relevant for England, as they could provide examples for improving the quality of education and assessment practices. For instance, the OECD report notes that the emphasis on competency-based learning and formative assessment aligns with current trends in England's education system. However, it also highlights that the decentralised system of assessment may pose challenges for transferability, as England's assessment system is more centrally regulated.

Finally, a report by Suto and Oates (2021)⁹ on high-stakes testing in high-performing education systems provided many useful insights. The report emphasises how national legislation, cultural circumstances, and public expectations influence high-stakes testing practises in Ontario, although arguably this applies to other jurisdictions. It suggests that it is essential to consider the policy context and cultural factors that shape the education system in England when assessing the transferability of these practices.

The report also highlights the need to understand the broader systemic elements that lead to high-stakes testing success in various countries (Suto & Oates, 2021). The systemic elements include teacher training and support, curriculum alignment, assessment design, and the utilisation of outcomes for accountability and school improvement. The report recommends that England investigate these systemic elements and apply them appropriately to its own environment.

In addition, the Suto and Oates (2021) report emphasises the importance of incorporating curriculum standards into assessment practises in order to assure the validity and reliability of high-stakes assessments. England could harmonise its curriculum and assessment frameworks to design clear objectives and guarantee that assessments successfully evaluate the desired learning outcomes.

Finally, the report underlines the need to involve many stakeholders in the design and implementation of high-stakes testing, such as teachers, school administrators, legislators, and the broader community (Suto & Oates, 2021). This was echoed in evidence identified in the research review: England may benefit from high-performing education systems' experiences in engaging stakeholders to develop support and ensure the effective implementation of any high-stakes testing reforms.

Overall, the transferability of national-level assessment policy and reforms between countries is complex, but there are clear insights to be gained.

Strengths and limitations of the review

As with any piece of research, there will be strengths and limitations of the work. In terms of **limitations**, the review question was very broad, which perhaps made it difficult to pin down key concepts. For example, a key aspect of research question 6 is to explore whether purposes were achieved, but the purposes were mostly not pre-defined—they were derived from what we saw in the evidence. This meant that study identification processes had to be

⁹ Suto, I., & Oates, T. (2021). High-stakes testing after basic secondary education: How and why is it done in high-performing education systems? Cambridge Assessment Research Report. Cambridge, UK

broad and there were inefficiencies in data extraction as we came to understand the data-driven themes.

Given the wide scope of the review, we did not identify a large number of studies in each jurisdiction—with no eligible studies for three jurisdictions. This inhibited the ability to take a comparative approach. We observed that there may be a difference between high and low stakes assessment, and the relative balance of formative and summative assessments, but not enough evidence to do a synthesis on these features.

In terms of our processes, we did not do an extensive critical appraisal of the studies. As a rapid review with limited resources, we took a necessarily pragmatic approach, but this may undermine the emphasis we place on studies. Ordinarily, the critical appraisal would be integrated into the synthesis more (Liabo, Gough, & Harden, 2017); here, we used it to identify major red flags in the set of studies.

Finally, while the team is highly skilled in terms of research methods, education research, and education policy, they had limited prior subject knowledge of secondary school summative assessments. Some aspects (e.g., developing the data extraction tool) may have been more efficient if a subject matter expert had contributed.

There were various strengths of the review. Importantly, we conducted a thorough and extensive review with good quality checks of processes and procedures. These included double screening, double data extraction, and decision making checked by multiple reviewers. The level of rigour went way beyond the typical expectations of a rapid review.

We were also able to adapt the synthesis methods to the evidence identified. For example, the novel introduction of the comparative synthesis added informational value above that proposed in the original review plan; this was done in response to observations of divergence and convergence in themes across the two types of evidence. This type of responsive synthesis was enabled by having a team of experienced researchers in systematic review methods and evidence synthesis.

Despite the paucity of studies from some jurisdictions, the review was able to address the research questions adequately. We managed to synthesise a lot of detail from the included studies, which meant that there was good depth of information for most aspects of the review. Related to this, we highlighted some important potential research gaps and drew together important points pertinent to developing future assessments, such as involving students in the design.

Deviations from proposal

Deviations from what was originally proposed allowed for more evidence to be included and for deeper synthesis to be conducted.

- We expanded research question 6 to remove the restriction to quantitative evidence.
- We widened student impacts to include stress and anxiety.
- We distinguished between assessment *purposes* and *uses*.
- We introduced a comparative synthesis, to investigate how the policy sources and research papers aligned.

Conclusion

We conducted a rapid systematic evidence synthesis of 21 studies on summative assessment in general lower and upper secondary education (ages 14-18/19) in 10 jurisdictions (England, Estonia, Finland, Hong Kong, Japan, the Netherlands, New Zealand, Ontario - Canada, Poland, and Republic of Ireland). We found that summative assessment is

a highly contentious topic, with debates on diverse aspects, including potential harms to students; issues around equity, fairness, and disadvantage; the form and function of the assessments; and the coverage of knowledge and skills. There is no one-size-fits-all approach to summative assessment, but a student-centred approach that fully integrates student and teacher voice and that takes into account the various layers of the system in which it operates is recommended for future policy reform.

References included in the review (n = 21)

Baird Jo-Anne, Caro Daniel H and Hopfenbeck Therese N (2016) Student perceptions of predictability of examination requirements and relationship with outcomes in high-stakes tests in Ireland *Irish Educational Studies*, 35(4), pp361-379.

Baird Jo-Anne and Ofqual (2019) *Examination Reform: Impact of Linear and Modular Examinations at GCSE*. Available at: https://www.educationox.ac.uk/wp-content/uploads/2019/04/Modular_linear_report_final.pdf.

Banks Joanne and Smyth Emer (2015) 'Your whole life depends on it': Academic stress and high-stakes testing in Ireland *Journal of Youth Studies*, 18(5), pp598-616.

Benton Tom (2014) Examining the Impact of Entry Level Qualifications on Educational Aspirations *Educational Research*, 56(3), pp259-276.

Brown Kerry and Woods Kevin (2022) Thirty years of GCSE: A review of student views and experiences *Assessment in Education: Principles and Policy & Practice*, 29(1), pp51-76.

Cadwallader Stuart (2013) Is the English Baccalaureate the most appropriate academic core? Subject choice and attainment at GCSE and A-level, pp16.

Carroll Matthew, Benton Tom and Cambridge Assessment United Kingdom; (2018) *The Link between Subject Choices and Achievement at GCSE and Performance in PISA 2015*: Cambridge Assessment, pp Available at: <https://search.ubscohost.com/login.aspx?direct=true&db=eric&AN=ED600647&site=ehost-live>.

Elwood Jannette, Hopfenbeck Therese and Baird Jo-Anne (2017) Predictability in high-stakes examinations: students' perspectives on a perennial assessment dilemma *Research Papers in Education*, 32(1), pp1-17.

Hargreaves A (2020) Large-scale assessments and their effects: The case of mid-stakes tests in Ontario *Journal of Educational Change*, 21(3), pp393-420.

Hepworth Neal, Galvis Maria and Gambhir Geeta ; Sizmur Juliet ; National Foundation for Educational Research (NFER) (United Kingdom; (2021) Using PISA 2018 to Inform Policy: Learning from the Republic of Ireland *National Foundation for Educational Research*.

Jerrim John (2022) The power of positive emotions? The link between young people's positive and negative affect and performance in high-stakes examinations *Assessment in Education: Principles and Policy & Practice*, 29(3), pp310-331.

Jerrim, J. (2022). 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>

Lessof Carli, Ross Andy and Brind Richard ; Harding Carrie ; Bell Emily ; (2018) Understanding KS4 attainment and progress: evidence from LSYPE2 , pp136.

Machin Stephen (2018) *Entry Through the Narrow Door: the Costs of Just Failing High Stakes Exams*. Available at: <https://cverlse.ac.uk/textonly/cver/pubs/cverdp014pdf/files/20354/cverdp014.pdf>.

McCoy Selina, Byrne Delma and O'Sullivan Jessica ; Smyth Emer. (2018) *The early impact of the revised Leaving Certificate grading scheme on student perceptions and behaviour*. Available at: https://www.esri.ie/system/files/publications/RS85_online1_0.pdf

Nieminen Juuso Henrik and Atjonen Päivi (2022) The assessment culture of mathematics in Finland: a student perspective *Research in Mathematics Education*, 0(0), pp1-20.

Ofqual (2012) *International Comparisons in Senior Secondary Assessment Full Report*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/372211/2012-06-12-international-comparisons-in-senior-secondary-assessment.pdf
[files/20360/2012-06-12-international-comparisons-in-senior-secondary-assessment.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/20360/2012-06-12-international-comparisons-in-senior-secondary-assessment.pdf)

Putwain David W, Daly Anthony L; Chamberlain Suzanne and Sadreddini Shireen. (2016) 'Sink or Swim': Buoyancy and Coping in the Cognitive Test Anxiety--Academic Performance Relationship. *Educational Psychology*, 36(10), pp1807-1825.

Smyth Emer and Banks Joanne (2012). High stakes testing and student perspectives on teaching and learning in the Republic of Ireland *Educational Assessment and Evaluation and Accountability*, 24(4), pp283-306.

Tong Siu Yin Annie and Adamson Bob (2015) Student voices in school-based assessment *Australian Journal of Teacher Education*, 40(2), pp15-28.

Zhao Anran, Brown Gavin T L and Meissel Kane (2022) New Zealand students' test-taking motivation: an experimental study examining the effects of stakes *Assessment in Education: Principles and Policy & Practice*, 29(4), pp397-421.

Appendix A. NFER Scoping study

Click on the image below to open the document.



Confidential

Report The NFER Research Programme

Education Policy Knowledge Centre

National Foundation for Educational Research (NFER)



Appendix B. Methods for the review of policy sources

We only considered evidence identified by NFER from their searching of key databases. NFER handed over an excel spreadsheet with a list of 326 sources across the 10 jurisdictions. The number of documents varied across the jurisdictions, ranging from 13 to 49. The sources were listed as web links only and there was no indication from the spreadsheet about the length or coverage of the included documents. Some sources were highlighted to potentially identify them as priority sources to address the research questions. However, a scoping exercise of one jurisdiction revealed that all sources listed under each jurisdiction would need to be considered in our synthesis.

These steps set out our process for reviewing the NFER sources for this phase of our study. We:

- Identified and described the sources of information to be used, such as government reports, private data, or Wikipedia articles.
- Reviewed the information from these sources and provide page numbers or title names, as well as assign a quality indicator based on factors such as credibility, bias, and relevance
- Reviewed the unhighlighted sources using the same process as in step 2.
- Identified and reconciled any discrepancies in the information extracted from both highlighted and un-highlighted sources, considering the nature of the source and the quality indicator assigned.
- Populated the jurisdiction brief template with these summaries.
- Compiled and included a list of references.
- Used the above to produce a report for each jurisdiction. The briefs are a synthesised summary of the information for each research question.
- Checked the draft report for each jurisdiction for coherence and proofread for errors
- A second researcher reviewed the draft report for each jurisdiction for quality.
- Sent a sample of the draft reports for each jurisdiction to the funder for checking and approval.

Each jurisdiction brief had the same structure, designed to address each of the research questions. The structure consisted of:

- **Background information** Here we provided a brief overview of the current summative assessment system(s) in place and discussed the specific aspects being explored, including the ages/stages at which assessments take place, forms of assessment (teacher assessment, examination, coursework), mode of assessment (paper-based, electronic), standardisation, length of time the system has been in place
- **A discussion of the stated purposes of the summative assessments** (including benefits) and how the outcomes are used (grading and sorting; certification; institutional accountability), reliability and validity, and the potential impact of publishing results.
- **A discussion of the national-level assessment policy process** Here we included a brief discussion on the key stakeholders involved in the policy process, including government officials, educators, and parents. Here we discussed the process through which national-level assessment policy is made, including who specifies what needs

to be assessed and how this is determined We also identified any regulatory authorities or awarding bodies that play a role in specifying minimum content.

- A summary of how stakeholders, including educators and parents are involved in the evaluation process We discussed the mechanisms through which the current summative assessment system is evaluated, including any feedback loops or continuous improvement processes.
- A discussion of any noted challenges in evaluating the summative assessment system, and discussed the contextual factors that have driven national-level assessment policy and/or any reforms over the last 10 years Part of this discussion was to try and assess what these factors mean for the transferability of the current system to the context in England and to highlight any obvious implications for policymakers and other stakeholders
- A summary of the findings of any research studies identified in the research review that related to the particular jurisdiction.

Appendix C. Data extraction tool

ASSESSMENT CHARACTERISTICS

- General context: Name or brief description of summative assessment system(s) as discussed in the paper
- Evaluation: Name or brief description of aspect of summative assessment system(s) evaluated in the paper
- Policy's targeted age/stage of education (add detail)
- Form of assessment
 - Teacher assessment
 - Examination
 - Coursework
 - Other (please add)
 - Unclear
- Mode of assessment
 - Paper-based
 - Electronic
 - Other (please add)
 - Unclear
- Is it standardised?
 - Yes
 - No
 - Unclear
- What are the stated PURPOSES of the summative assessment? That is, for what explicit PURPOSES were the outcomes of summative assessments designed to be used (eg grading and sorting; certification; institutional accountability)
 - 1 To progress through levels of education (Combined theme) Merger of three themes of progressing through (1) higher levels of secondary education, (2) vocational/technical education, and (3) tertiary/higher education
 - To determine pupil access to tertiary/university/higher education
 - To progress to higher levels of secondary school education (e.g., GCSE's to A-levels)
 - To determine access to vocational/technical education, internships, etc
 - 2 To assess whether learning objectives have been achieved, or assess pupil knowledge/skill acquisition
 - To allow employers to differentiate between candidates
 - 3 For QA/ accountability of school or system (Combined theme), Merged two themes of assessing school or system level QA or accountability, informing parents
 - To assess school- or system-level performance for quality assurance or accountability
 - To assess school-level performance to inform parents/carers decisions about school applications
 - 4 To enable international comparisons (e.g., PISA)
 - 5 To ensure fairness or equality in assessment
 - Other (please add)
 - Unclear
- What are the stated USES of the summative assessment? That is, for what explicit USES of the outcomes of summative assessments (eg grading and sorting;

certification; institutional accountability), whether they were intended in the original policy development or not

- To determine pupil access to tertiary/university/higher education
- To determine access to vocational/technical education, internships, etc
- To progress to higher levels of secondary school education (e.g., GCSE's to A-levels)
- To assess whether learning objectives have been achieved, or assess student progress (i.e., pupil knowledge/skill acquisition)
- To allow employers to differentiate between candidates
- To assess school- or system-level performance for quality assurance or accountability
- Incentivising teachers and/or school managers
- To assess school-level performance to inform parents/carers decisions about school applications
- To enable international comparisons (e.g., PISA)
- Other (please add)
- Unclear

STUDY CHARACTERISTICS

- Year the study commenced (or publication year if unclear)
- Jurisdiction (tick all that apply)
 - England
 - Estonia
 - Finland
 - Hong Kong
 - Japan
 - The Netherlands
 - New Zealand
 - Ontario (Canada)
 - Poland
 - Republic of Ireland
 - Includes other jurisdictions
- Coverage of the evaluation (sampling)
 - Nationwide
 - Region/s
 - Systematic review (coverage may vary by included study)
- Sample
 - Age as reported by authors or as per SR inclusion criteria (add info) As reported by the authors (range, mean, other) or based on a systematic review inclusion criterion
 - Sample size in primary study evaluation (add info)
 - Number of participants/respondents in the primary study Do not use for systematic reviews
 - Sample size in systematic review (number of studies; add info) Number of studies included in the review Do not use if it's a primary study
 - Sample characteristics (tick all that apply) If there are separate analyses for universal and subgroups, you can tick both
 - Universal (general population)
 - Specific characteristics (e.g., SEND) - (add info)
 - Unclear sample characteristics

RESULTS OF THE STUDY

- Did the study attempt to evaluate the stated purpose/s of the policy in some way?
 - YES, study explicitly evaluated the stated purpose/s of the assessment policy
 - If YES, what did they find in relation to the stated purpose/s? ADD info
What does this research tell us about the extent to which the assessment systems meet the purposes stated in above?
 - NO, the study evaluated another aspect of assessment not related to the stated purpose In other words, stated purpose of assessment not addressed in the evaluation
 - If NO, move on to the next question
- Attainment
 - Reported (add study conclusion)
 - Not reported
- Students' motivation to learn
 - Reported (add study conclusion)
 - Not reported
- Stress/anxiety
 - Reported (add study conclusion)
 - Not reported

DEBATES ABOUT CURRICULUM ASSESSMENT (ADD THEMES)

- Implications for discussions about summative assessment Can include discussion of transferability of findings to other contexts
- Potential harms to students (e.g., stress)
- Implications for teacher practice
- Implications for future assessment
- Technical issues (e.g., reliability, validity)
- Predictability of content; teaching to the test
- Relevance of disadvantage
- High stakes versus low stakes
- Student voice in policy reform
- Breadth versus depth of subject matter coverage
- Policy reform

CRITICAL APPRAISAL

- Did the research have an advisory panel and/or ethics approval?
 - Yes
 - No
 - Unclear
 - Second reviewer query
- Richness of study (amount of content that speaks to the review's questions)
 - High (rich coverage of our review aims)
 - Low (limited scope related to our review question)
 - Second reviewer query
- Are the methods clearly stated?
 - Yes - transparency of methods
 - No - important aspects of methods not reported or unclear
 - Second reviewer query
- Are serious limitations raised by the study author/s?

- Yes - serious limitations mentioned by the authors
 - Yes - limitations noted but not considered serious by the authors
 - Limitations not adequately discussed
 - Second reviewer query
- Reviewer comments on limitations
 - Add any reviewer concerns about methods here
 - No comments to make

Appendix D. Study findings about progression to the next stage of education

Author	Jurisdiction	Assessment	Age range covered by assessment	Findings about progression
Baird 2016	Republic of Ireland	Irish Leaving certificate	Median age of examinees = 18	The qualification reported in this paper is the Irish Leaving certificate which is taken when young people are around age 18, so progression here is into higher education or employment. Student participants reported the tests undertaken for the Irish Leaving certificate (regardless of subject) did not test the right kind of learning. Only one third of respondents said they would be able to use what they had learned in the future.
Baird 2019	England	General Certificate of Secondary Education (GCSE)	14-16	Progression in this paper is between GCSE and A level qualifications. Teachers interviewed for the study reported that content of the assessment was correctly geared towards a route into A levels. However, there were concerns that the assessment 'would potentially leave more of the lower attaining students behind'. There was also expressed concern that students who were more interested in a vocational route would lose out from the increased emphasis on the more traditional curriculum. The study reported not finding any evidence of effects of the format of the assessment (GCSE) upon progression to A level While teachers considered a 'modular format' would better support outcomes (and therefore progression), especially of 'lower attainment students', this was not supported by the study's findings on attainment. Some teachers

Author	Jurisdiction	Assessment	Age range covered by assessment	Findings about progression
				maintained that linear assessments were the gold standard for student progression.
Caroll 2018	England	PISA and GCSE	14-16	Only the purpose was mentioned. No results about progression.
Elwood 2017	Republic of Ireland	Irish Leaving certificate	16-18	This paper discusses the Irish leaving certificate. As this is taken by age 18, progression here is onto higher education or employment. The authors comment that the assessment helps prepare young people for the future and 'the world outside of school' Pupils said they took the assessment to 'optimise their performance' in the assessment (in order to stand a better chance of progression to higher levels of qualification). One student participant commented that "I'm more kind of learning how to get as many points as possible, not trying to learn to get knowledge out of the subject ... all I want to do is get as many points as I can to get to a college course" (p. 14).
Machin 2018	England	General Certificate of Secondary Education (GCSE)	16	Progression in this paper is between GCSE and A level qualifications. Certain minimum grades were reported as being needed for progression, and failure to achieve these result in students having to re-sit assessments. Page 10 for example says: "from 2015 onwards, it has been compulsory for students who do not achieve a C in English or maths to re-sit the exam over the next year (which is typically in a college of further education, where such students will be most likely enrolled in some form of vocational education)". The thresholds were reported to have serious implications for progression, with "about 9 percent of students have dropped out of any form of education by the age of 18 (rising to 11 percent of students within +/- 10 marks of the Grade C threshold)"

Author	Jurisdiction	Assessment	Age range covered by assessment	Findings about progression
				(p. 23). The results suggest "about 90% of people (in the range of marks from grades C to D) manage to start a high-level qualification within this time and thus it is not a very high yard-stick of achievement Yet, just failing to get a C grade manifestly has a huge effect on the probability of getting back on track within 3 years" (P. 24). The study also found that being on the borderline of achieving grade C affected the probability of students enrolling in tertiary education: "Marginally achieving a C grade increases the probability of commencing tertiary education by 25 to 4 percentage points in a context where about 27 percent of this sample have started tertiary education by this age (205 percent for those within 10 marks of the C threshold)" (P. 24).
McCoy 2018	Republic of Ireland	Irish Leaving certificate	12 to 15	Only the purpose was mentioned. No results about progression.
Ofqual 2012	England; Other countries (Finland, Hong Kong, Netherlands, New Zealand, Republic of Ireland)	Advanced Level General Certificates of Education (A levels); range of assessments	18	Progression in this paper for England is between A level qualification and higher education or employment pathways. The report findings state that the assessments being reviewed offer students "a broader range [that] allows a student greater flexibility in the pathways that they can study in higher education" (p. 13) It was noted for example that "A levels offer one of the broadest programmes of study seen" across a range of jurisdictions analysed (p. 13). A levels were noted to offer breadth as well as depth, particularly of science subjects such as Biology. However, A levels were noted to be more focused on reading and interpretation of text, rather than emphasis on "practical skills", which could also be of benefit to further education.

Appendix E. Study findings about student learning

Author	Jurisdiction	Assessment	Age range covered by assessment	Findings about student learning
Benton 2014	England	Entry-Level Qualifications (ELQs)	16	ELQs are ‘aimed at learners who may not be ready for a GCSE, or who will benefit from reinforcing their basic ... skills while studying towards the GCSE’ (p. 3). The study aimed to assess the extent to which taking ELQs is associated with increased aspirations for future study and learning. The reports main finding was that the ELQ (Entry Level Qualifications) was associated with increased chance of pupils staying in education post-16. However, ELQs were not seen as beneficial for entering HE (Higher Education) in the short-term. The report noted that “two thirds of these pupils were at a level of attainment where taking an ELQ may have been beneficial” for learning outcomes (p. 15).
Brown 2022	England	General Certificate of Secondary Education (GCSE)	16	The overall study aimed to investigate young people’s views and experiences of GCSE study and assessment. In terms of use of GCSEs (General Certificate of Secondary Education) to meet learning objectives, students expressed some frustration at not being “shown properly how to actually do the coursework and the exam” (p. 17). They also noted a lack of social equity in being supported with undertaking

Author	Jurisdiction	Assessment	Age range covered by assessment	Findings about student learning
				and preparing for exams, commenting that some pupils benefited unfairly from parents who could afford to pay for private tuition support.
Carroll 2018	England	PISA GCSE	16	In this study, the PISA (Programme for International Student Assessment) was noted to be “designed to measure the extent to which pupils can apply particular areas of knowledge to solve problems within real world situations”. It was also noted that “GCSEs are intended to test individual pupils’ knowledge of and ability to apply specific content as defined in the qualification specification” (p3). PISA was noted to measure different things to GCSEs in terms of student learning GCSEs were noted to better capture individual student learning and to measure “the extent to which pupils can apply particular areas of knowledge” (p. 3).
Elwood 2017	Republic of Ireland	Irish Leaving Certificate	16-18	The report aimed to evaluate students' reflections about the definition of predictability within the context of the Irish Leaving Certificate. Having predictability in the assessment was argued to focus student learning of the content and skills required across a broad variety of subjects studied. The report comments that while “the aims and objectives of the system are to engender positive learning outcomes and prepare young people for the world outside of school, the reality is that young people are consumed with optimising their performance on these examinations and see this as the main goal of their schooling” (p. 15). The paper argues that “neglecting to include students’ reflections on these perennial assessment dilemmas fails to see them in their totality and means we miss valuable insight into how

Author	Jurisdiction	Assessment	Age range covered by assessment	Findings about student learning
				assessment problems play out in reality for test-takers” (p. 15). The paper calls for students to be included in assessment reforms since they are knowledgeable about what methods work best for them.
Nieminen 2022	Finland	Assessment for Learning	12 and 15	The ‘Assessment for Learning’ is noted to be a ‘low stakes’ measure for developmental evaluation purposes. This study aimed to evaluate student perspectives of how mathematics is assessed in Finland and compared a variety of assessment methods at testing student knowledge. The study comments that “students perceive the assessment culture of mathematics in Finland as exam-driven: a closed-book exam appears to be the “default practice” (p. 14). The paper discusses how this perception contradicts the national guidelines, “in which students’ active role and participation in mathematics assessment” is encouraged (p. 14). An explanation for this perception provided by the authors is that “exams are deemed suitable for assessing mathematical knowledge” (p. 14). Another is that the findings reflect “the complexity of implementing formative assessment” and the “broader educational and pedagogical studies of their teachers” (p. 14). The authors call for “greater consideration of the students’ perspective” within assessment development (p. 17).
Ofqual 2012	England; Other countries (Finland, Hong Kong, Netherlands, New Zealand,	Advanced Level General Certificates of Education (A levels); range of assessments	18	This study reviewed four subjects (Mathematics, Chemistry, English, and History) in 5 jurisdictions (named in next column) in comparison with A levels in each of those subjects in England In the context of this theme, the study evaluated how these assessments identified young people suitable for higher education.

Author	Jurisdiction	Assessment	Age range covered by assessment	Findings about student learning
	Republic of Ireland)			The report found that “The courses which provide the best foundation for study at a UK (United Kingdom) university include a focus on genre-based study over single text study, testing of both content and skills, and where components are organised according to a coherent set of theoretical principles” (p14). Some points specific to the individual subjects were also noted for benefiting student learning. For example, using coursework and personal study in A levels was particularly beneficial to historical studies, and A level Further Mathematics was found to be the broadest and deepest of the various qualification reviewed for this study (p. 14).
Tong 2015	Hong Kong	School-based Assessment (SBA), an integral part of the English Language assessment for the Hong Kong Diploma of Secondary Education (HKDSE)	15	School-based Assessment (SBA) in Honk Kong is an English language-based assessment that includes formative and summative elements. Completion of it contributes to the Hong Kong Diploma of Secondary Education (HKDSE) (which was implemented from 2012). The report found that the SBA stimulated “higher-level thinking and learning, and that helped structure the learning process for work that was carried out in groups” (p. 10). However, “students were not satisfied with the general feedback given as it was not particularly relevant to one specific task,” and many students did not know how to apply the feedback given (p. 10) Students were reported to be “sceptical about the capacity of SBA to perform the functions ascribed to it, such as advancing their learning and language skills (items 1, 5 and 6), or measuring their knowledge and skills (items 2, 3 and 12)” (p. 7). The report concluded that “Student voices are to be valued because their perceptions influence how they learn and affect the quality of their learning” (p. 26). This study also noted the “challenge facing curriculum developers is to find ways to incorporate a range of assessments—formal

Author	Jurisdiction	Assessment	Age range covered by assessment	Findings about student learning
				and informal; standardised and school-based—in a balanced and principled manner” (p. 26).

Appendix F. Study findings about equality and fairness within the education system

Author	Jurisdiction	Assessment	Age range covered by the assessment	Findings about equality and fairness within the education system
Benton (2014)	England	Entry-Level Qualifications (ELQs)	16	For pupils with low levels of prior attainment, ELQs are associated with significantly increased chances of pupils stating that they wish to stay in education post-16 and that they are intending to apply to university in the future (p. 267). As well as being associated with aspirations for future study, ELQs were associated with low achieving students staying in education (ibid). However, ELQs were not considered good preparation for higher education, even for low achieving students (p. 269).
Brown (2022)	England	General Certificate of Secondary Education (GCSE)	16	This study was a synthesis of evidence from published literature in 22 included studies. Several factors were cited as important in considering fairness and equality within GCSEs.

Author	Jurisdiction	Assessment	Age range covered by the assessment	Findings about equality and fairness within the education system
				For example, having a choice of assessment (e.g., exams and coursework) was seen as important (p. 65). Some students raised concerns about teacher bias in teacher assessments (ibid), and ‘gender stereotyping’ (ibid). For example, girls found it easier to work hard for GCSEs without any negative effect upon peer acceptance (ibid). “Young people linked a perceived lack of fairness and equality within their experiences of GCSE with a detrimental impact on enjoyment, motivation and ultimately attainment” (ibid). Some young people expressed concerns that GCSE reforms have led of differences in assessment requirements across different assessment contexts, such as the countries of the UK (ibid).
Cadwallader (2013)	England	The English Baccalaureate (EBacc)	16	The core subjects studied for the EBacc were questioned in this study In particular, the marginalisation of vocational subjects and the arts. The current core subjects were perceived to be potentially disadvantaging “pupils who are not academically inclined” (p. 2-3).

Author	Jurisdiction	Assessment	Age range covered by the assessment	Findings about equality and fairness within the education system
				The authors cite a House of Commons report (2011) which said that “academic subjects are not the only path to a successful future, and all young people, regardless of their background, must continue to have opportunities to study the subjects in which they are likely to be most successful, and which pupils, parents and schools think will serve them best”, p. 4).
McCoy (2018)	Republic of Ireland	Leaving Certificate (LC)	18	Grading reforms which had led to wider grade bands were perceived as being unfavourable to students who received grades at the top of the grade band (p. 32). Furthermore, placing specific emphasis on specific subjects (i.e., higher level subjects such as mathematics) was also perceived as being disadvantageous to students who may not be academically inclined in those subject areas P41).

Appendix G. Study findings about student attainment

Author	Findings	Jurisdiction	Assessment	Age range covered by assessment
Baird 2016	Students thought the assessment did not cover the right kinds of learning (p27) Of particular concern was that students felt examinations placed too much pressure on memory. “Students who believed narrowing of the curriculum was a good tactic tended to perform worse in the English and Biology examinations” (p. 28). “Students who perceived that preparing for the examination was a valuable learning experience tended to score higher in the English and Biology examinations” (p. 28).	England	Entry-Level Qualifications (ELQs)	16
Baird 2019	“We found no educationally significant effect of examination structure on grade outcomes in English, mathematics or science. Nor did we find any statistical evidence of effects of structure upon progression to A level So there was no evidence to suggest that either modular or linear GCSEs lead to better educational outcomes Importantly, equity gaps in gender, socioeconomic status and school type were not affected by structure once prior attainment was taken into account. While some teachers believed that certain groups of students would perform better in a modular system, this view was	England	General Certificate of Secondary Education (GCSE)	16

Author	Findings	Jurisdiction	Assessment	Age range covered by assessment
	not supported by analysis of outcome data, where such beliefs could be investigated" (p. 140). "Findings do not support claims that modular or linear exams tend to favour male or female students or influence low and high SES students differently As expected, high SES students performed better than low SES students in GCSEs and female students performed better than male students in GCSE English. However, GCSE gaps related to gender and SES remained relatively unchanged for linear and modular examinations Results relating to GCSE gaps between state and independent schools were less clear There was weak evidence that state schools performed better in modular than linear exams compared to independent schools" (p. 136).			
Benton 2014	In relation to the main impact of ELQs, the results report that "no statistically significant effects were found relating to this qualification type", indicating that "overall, there is no evidence of these qualifications having either a positive or a negative effect on the future aspirations and destinations of pupils" (p. 10). Comparisons between low and high achieving students on this qualification are noted to be flawed because "ELQs were never designed or intended to be taken by high attaining pupils" (p. 16).	England	Entry-Level Qualifications (ELQs)	16

Author	Findings	Jurisdiction	Assessment	Age range covered by assessment
Brown 2022	Teachers were reported to be “an important source of educational and emotional support for motivation and, ultimately, attainment at GCSE” (p. 18). Some young people expressed concerns around GCSE reforms and how this was affecting attainment between different groups of young people (p. 17). Some young people linked successful performance on exams to having private tutors (p. 17).	England	General Certificate of Secondary Education (GCSE)	16
Cadwallader 2013	The study reported a strong link between the EBacc and prior attainment (p. 12). They argue that student choice of subjects in the EBacc leads to better attainment outcomes, with those subjects of interest to individual students resulting in better grades (p. 12). The authors found “some evidence for a small but statistically significant effect of studying the EBacc being carried over to A-level, whereby those with the EBacc scored about half a grade more on their total A-level score” (p. 13). The study suggests “the combination of subjects which is at the core of the EBacc may be associated with higher overall academic performance” (p. 13). Furthermore, students who studied the constellation of subjects included in the EBacc “attain, on average,	England	The English Baccalaureate (EBacc)	16

Author	Findings	Jurisdiction	Assessment	Age range covered by assessment
	just under half of a grade more (451) at GCSE” (p. 8). Additionally, “Whether or not a pupil took GCSEs that could have led to the EBacc does seem to influence A- level performance: those who took the EBacc achieve just over half a grade (509) more on their A-level score” (p. 9). However, the authors caution that the EBacc only explains a very small amount of the variance in GCSE and A level attainment overall (p. 8 and 9). Furthermore, given that those pupils who take A-levels are “a self-selecting group of relatively high achievers”, they say that the results of effect on A level is not particularly surprising (p. 13).			
Carroll 2018	The authors report that “our analysis suggests that roughly a tenth of students averaging at grade C at GCSE (ie of middling ability in GCSEs, see Gill, 2017) are amongst the top fifth of students when it comes to collaborative problem solving (see main report). Even amongst pupils with an average GCSE grade of D, one in twenty will be amongst the top fifth of performers in collaborative problem solving” (p. 5). They also state that “Even after accounting for differences in achievement at GCSE, in every domain, those studying separate sciences were around 20 PISA points ahead of those studying combined science – a difference equivalent to roughly 6 months progress” (p. 6). Also, that “the differences in medians between	England	PISA GCSE	16

Author	Findings	Jurisdiction	Assessment	Age range covered by assessment
	those taking separate and combined sciences was around 40 PISA points amongst those with grade D, around 25 points amongst those with grade A, and 10 points amongst those with grade A*" (p. 6).			
Hepworth 2021	This report's context is that the Republic of Ireland achieved the highest PISA score for Reading up to 2009. The report looks into the factors why this has been the case It concludes that "the most notable differences being pupils in the Republic of Ireland report a greater enjoyment of reading and a greater number of cultural possessions in their home" (p. 62). Additionally, "Targeted support through DEIS and the NLNS was judged to have played a role in this success by evaluations and the experts in this study" (p. 64).	England, Republic of Ireland	PISA GCSE	15
Jerrim 2022a	The results find a significant association between emotion and attainment at GCSE level However, the authors comment the effect size for this is 'trivially small' (p. 16). The authors find no association between "fear of failure and GCSE grades" (p. 17).	England	General Certificate of Secondary Education (GCSE)	16
Jerrim 2022b	The authors find "little evidence of a clear link between test anxiety and GCSE performance" (p. 14), and this also applies to sub-groups such as SES.	England	General Certificate of Secondary Education (GCSE)	16
Lessof 2018	The authors comment that there is evidence that "many aspects of disadvantage were associated with lower attainment, even after controlling for other factors " (p. 12). Furthermore, "disadvantage is likely to have a layering effect – the more elements of disadvantage a young person faces, the lower their attainment is likely to be	England	General Certificate of Secondary Education (GCSE)	16

Author	Findings	Jurisdiction	Assessment	Age range covered by assessment
	Elements of disadvantage which were found to be associated with lower attainment, even after controlling for all the other modelled factors" (p. 12). The authors also found that several factors were related to low attainment (p. 13): • "Young people who suffered from cyber-bullying, all other factors being equal, were found to have lower attainment than those who did not" • "Young people with challenging parental relationships showed lower attainment regardless of their background – those whose parents reported that they argued more than once a week had lower attainment than those who argued less often" • Sleeping for nine and a half hours or more per night (that is, more than the optimal amount of sleep) was associated with lower attainment (even after factors such as psychological distress and the other modelled variables had been controlled for)" • "Truanting for several days at a time (or longer) corresponded to markedly lower attainment" • "Young people who usually did all of their homework had higher attainment than those who usually did most of their homework and, particularly, those who usually only did some or none of it".			

Author	Findings	Jurisdiction	Assessment	Age range covered by assessment
	Another important factor is ethnicity The authors say that “after controlling for all the other modelled factors, the attainment of young people from Black African and Black Caribbean backgrounds was slightly lower than that of other groups, a finding that would merit further investigation” (p. 14). However, some factors were not as important For example, the authors argue that “While the attainment of boys was lower than that of girls in absolute terms, this gap was no longer significant when other factors were controlled for” (p. 13). They found that other related factors were more important than gender per se. For example: “The higher incidence of Special Educational Needs (SEN) and disabilities which affect schooling among boys” (p. 13).			
Machin 2018	The authors found that “higher tier students are much better performing than lower tier students” (p. 13). They conclude from their analysis of GCSE results that getting a grade C is within the English education system is very important to future success (p. 16) For example, they say that “just failing to get a C grade manifestly has a huge effect on the probability of getting back on track within 3 years” (p. 24). The authors also evaluated students who got close to the threshold and contested the grading. They found that “a high proportion of students who request a re- mark do not actually cross the relevant	England	General Certificate of Secondary Education (GCSE)	16

Author	Findings	Jurisdiction	Assessment	Age range covered by assessment
	threshold, and that crossing it is only likely for those students that originally scored a mark very close to the threshold" (p. 16).			
McCoy 2018	The Irish Leaving Certificate is used to progress students to higher education, and new grading bands were applied to the LC (prior to this research study) to “encourage the take-up of Higher Level subjects at Leaving Certificate” (p. 32) The grading separates subjects into ‘ordinary’ and ‘higher’. “Increased higher-level take-up has not translated into increased attainment, as there has been a significant drop in grades, all else being equal, for mathematics and Irish at higher level and Irish at ordinary level” (p. ii). The authors comment that “The models confirm that the drop in average grades is significant, all else being equal, for mathematics and Irish at higher level and for Irish at ordinary level” (p. 68). “Both higher- and lower-performing students felt that the gap between the points awarded for higher and ordinary papers is too wide and that ordinary-level points do not fairly reflect the workload and effort involved” (p. ii). The authors say policymakers need to look LC as it is concerning that their findings show students struggle to meet the requirements of the higher-level courses (p. 68). They also suggest the language of ‘higher’ and ‘ordinary’ levels for subjects being studied	Republic of Ireland	Irish Leaving Certificate	18

Author	Findings	Jurisdiction	Assessment	Age range covered by assessment
	is unhelpful and potentially demoralising for students studying the 'ordinary' level (p. 68).			
Putwain 2016	This study found that "the relationship between worry and GCSE score was partially mediated by task-focus and orientation" (p. 13). Academic buoyancy was found to moderate the relationship between the worry and task-focus and orientation (p. 13). In particular, "the relationship between worry and task-focus and orientation was stronger for low academic buoyancy" (p. 13) Academic buoyancy was defined as "withstanding and successfully responding to routine school challenges and setbacks".	England	General Certificate of Secondary Education (GCSE)	16

Appendix H. Study findings about student motivation

Author	Findings	Jurisdiction	Assessment	Age range covered by assessment
Baird 2019	This study showed mixed opinions by teachers about the impact of GCSE exams on motivation. For example, some teachers who were interviewed for the study considered that GCSEs were good at motivating students who had done well on early modules (p. 138). These teachers said doing well in exams raised student expectations (p. 86). Teachers who were more positive about the assessment said that “regular stepping stones of modular exams could keep students motivated” (p. 87). “Setting targets and getting feedback on those in the short-term” was also suggested by the authors to be motivational (p. 19). The authors said a head teacher commented “it was important to turn students into ‘resilient learners’ and ‘it’s okay to fail’” (p. 88). However, the authors note that an early module failure in exams could also have negative consequences. For example, it could provide a	England	General Certificate of Secondary Education (GCSE)	16

Author	Findings	Jurisdiction	Assessment	Age range covered by assessment
	'reality check' to prevent them falling into 'denial' or, on the other hand, it could also be a 'demotivator' (p. 87). Additionally, motivation was perceived by teachers to less good when students worked in groups, where it was observed to have been "demotivating" (p. 138). In such cases, "the lack of opportunity to 'put it right' in the case of a failure" was also identified as a demotivating factor by teachers (p. 87). Also, that students overall lacked motivation "due to the 'hardness' of the new examinations" (p. 113). Of particular relevance to reform of the assessment, was that teachers raised concerns "that not having any equivalency between the results this year and previous years would result in not only demotivation of students, but of teachers as well" (p. 108).			
Brown 2022	Relationships with teachers was cited as an important factor in motivating students to study and providing emotional support (p. 18; p. 67). One student said for example "if I don't get on with the teacher then not a lot tends to get done" (ibid). Another said, "if I can get on with the teacher then that's like I usually want to work well for them" (ibid).	England	General Certificate of Secondary Education (GCSE)	16

Author	Findings	Jurisdiction	Assessment	Age range covered by assessment
	The authors comment that students had both positive and negative feelings about GCSEs (p. 20). The authors say that while the needs of 'competence', 'autonomy' and 'relatedness' were met through doing GCSEs for some students, for others they were detrimental to motivation (ibid).			
Hepworth 2021	Nothing was explicitly written about this the theme of motivation However, motivation was alluded to in terms of factors that supported student enjoyment of subjects. For example, the authors said enjoyment of reading supported their attainment (p. 62) The authors also said that "integrated policy making which is co-created with teachers" was important for "keeping teachers on board and maintaining enthusiasm" (p. 59).	England, Republic of Ireland	PISA GCSE	15
McCoy 2018	The study found that the separation of grades into 'ordinary' and 'higher' levels impacted student motivation to study (p. 79), with low scoring being a demotivator (p. . iii). Teachers said students were only motivated to learn what they would be assessed on, and less interested in things outside of that (p. 82). There was a suggestion related to this, that students strategically choose subjects that are easier to pass (p. 83).	Republic of Ireland	Irish Leaving Certificate	18

Author	Findings	Jurisdiction	Assessment	Age range covered by assessment
Smyth 2022	Teachers own enthusiasm and love for the subject were seen as crucial factors in motivating student learning (p. 16). Students reported knowing when teachers loved their jobs and teaching “You know when you know that a teacher just, you know that they don’t want to be there and they’re just sitting there” (p. 16). The findings suggest that frequent test-based assessments were useful tools for keeping students motivated (p. 19). As was supported teacher learning, such as through revision sessions, which didn’t just leave the learning up to students (ibid).	Republic of Ireland	Irish Leaving Certificate	18
Zhao 2022	This study found that students’ perceptions of tests were a meaningful predictor of “test- taking motivation” (p. 15). The authors found that “students’ general belief about tests was a positive indicator of their perceived effort expenditure” (p. 2). In other words, how students perceived test taking affected their motivation to study, and consequently their attainment. Effort and motivation were found to be “notably lower” when students undertook national level assessment versus assessments that had a consequence for the individual (abstract). Students adjusted their perceptions of the test importance in accordance with the consequences attached to them (p. 410).	New Zealand	National Certificate of Educational Achievement (NCEA)	11-13

Appendix I. Study findings about student stress and anxiety

Author	Findings	Jurisdiction	Assessment	Age range covered by assessment
Baird 2019	Interviewed teachers raised “significant concern about the rising impact of the linear structure on students’ mental health” (p. 128).	England	General Certificate of Secondary Education (GCSE)	16
Banks 2015	Students reported stress related to the “importance of the impending exam for their future pathways” in the final year (p. 12). Female students were noted to feel more stressed than male students (p. 9). The authors say while most of this stress was “internally driven” pressure (ibid), teachers were noted to contribute to “fostering a ‘hot-house’ climate of exam preparation” (ibid). Additionally, the “schooling process” was noted to have an impact on stress levels (p. 2). The teacher-student relationship was cited as an important factor – “Students with more positive interactions with their teachers have lower stress levels, with higher stress levels among those who have experienced bullying from peers” (p. 2). Stress was also lower for students who were “facilitated in their choice of subjects” for their final years of study (ibid).	England	General Certificate of Secondary Education (GCSE)	16

Author	Findings	Jurisdiction	Assessment	Age range covered by assessment
Brown 2022	Students reported not enjoying their GCSEs because of the stress of the exams ('test anxiety') and fear of failure (p. 18). Students also reported concerns with "the need to manage simultaneous deadlines as well as wider commitments" (ibid). The relationship with teachers was also a cited factor in this study: "Some young people, even those with generally positive relationships with teachers, referred to teachers as 'stress amplifiers' who 'pressurise students'" (ibid).	England	General Certificate of Secondary Education (GCSE)	16
Elwood 2013	Students expressed concerns of qualifications being 'devalued' with assessment reforms (p. 10). They wanted results to be as valued as previous years as they were conscientious of the importance of this for them accessing higher education (ibid).	Republic of Ireland	Irish Leaving Certificate	16-18
Hargreaves 2020	Assessments were criticised for resulting in "ill-being" of students (p. 23). The study reported a strength of feeling among teachers that assessments were not neutral but "rather an intervention that can actively harm their students' well-being" (p. 21).	England, Republic of Ireland	PISA GCSE	15
Jerrim 2022	This study set out to examine test anxiety associated with GCSE examination. It found "very little difference in GCSE outcomes" between young people with different anxiety levels (p. 13). They conclude that "test anxiety is not related to the grades young people achieve in their GCSEs" (p. 16).	England	General Certificate of Secondary Education (GCSE)	16

Author	Findings	Jurisdiction	Assessment	Age range covered by assessment
Lessof 2018	This study, using data from the LSYPE, found “young people reporting particularly low levels of psychological distress or particularly high levels of psychological distress had lower attainment than those with moderate levels of psychological distress” (p. 13). They found that “young people who are “too laid back” or “too stressed” are both disadvantaged” (p. 87). However, they say further research is needed to confirm their findings, which argue “a little bit of stress is a good thing” (ibid).	England	General Certificate of Secondary Education (GCSE)	16
McCoy 2018	The authors say that “student stress remains a significant challenge for schools”, with head teachers reporting “an increase in emotional, psychological and mental health difficulties” (p. 100). Students spoke in the study about “constant levels of stress” (p. 102). The report goes on to say that “The large volume of information covered by students within subjects, along with the number of subjects, was identified as a precursor to the high stress levels” (p. 101). Teachers reported that student attendance was being affected by stress, with it manifesting differently for boys and girls (ibid).	Republic of Ireland	Irish Leaving Certificate	18
Nieminen 2022	In this study, “students generally reported relatively low levels of critical assessment pressure” (p. 14).	Finland	Assessment for Learning	12 and 15

Author	Findings	Jurisdiction	Assessment	Age range covered by assessment
Putwain 2016	The authors report that students' approach to coping with the pressure greatly affected the amount of test anxiety experienced. Academic buoyancy was found to moderate the relationship between the worry and task-focus and orientation (p. 13). Academic buoyancy was defined as "withstanding and successfully responding to routine school challenges and setbacks".	England	General Certificate of Secondary Education (GCSE)	16
Smyth 2022	Students, especially female students, reported high levels of test-related stress (p. 14). Students reported the pressure from teachers about the consequences of tests: "Teachers make out that if you fail the Leaving, it's the end of the world or something" (ibid). Students described how the preparation for the tests and an "escalation of demands" was affecting their enjoyment of the subject and their happiness, for example one said "you don't have a laugh anymore... We don't socialize either You stay in and study and not go anywhere and stuff like that and no one really goes out" (ibid). The authors comment that "Given the even higher stakes attached to Leaving Certificate performance, it is not surprising that these final year students report particularly high levels of pressure and stress" (p. 20).	Republic of Ireland	Irish Leaving Certificate	18

Author	Findings	Jurisdiction	Assessment	Age range covered by assessment
Zhao 2022	This study reports an increase in test anxiety related to “test stakes to the next level” (p. 13). Stress was adjusted according to the consequence of the test – those assessments which would have a more personal consequence, were associated with higher levels of test anxiety (ibid).	New Zealand	National Certificate of Educational Achievement (NCEA)	11-13

¹ At this stage, there is no high-stakes exam. Students are assessed through a variety of assessments throughout the academic year.

² The VMBO is a four-year programme that combines vocational training with theoretical education in languages, mathematics, history, arts, and sciences.

³ TSA = Territory-wide System Assessment.

⁴ This stage is not distinguished as part of what Ontario calls a 'continuous educational system' for ages 14/15 through to 18.

⁵ National Certificate of Educational Achievement.

⁶ Various types of upper secondary school: general secondary school, technical school, sectoral vocational school, and special school. The Matura is a high-stakes exam, Polish equivalent of the French Baccalaureate or English A-levels.

⁷ Upper stage: Operates from grades 9 to 12. The Ontario Secondary School Diploma (OSSD) includes a grade 9 mathematics assessment. The Ontario Secondary School Literacy Test (OSSLT) or the Ontario Secondary School Literacy Course (OSSLC), also in grade 10. A random selection of students is assessed for the Pan-Canadian Assessment Programme (PCAP), in grades 8 and 10. The Specialist High Skills Major (SHSM) enables learners to concentrate on a particular economic area while earning credits towards their Ontario Secondary School Diploma (grades 11 and 12).

⁸ There are currently no summative national examinations equivalent to the NAAA at the end of upper secondary school, although it was reported (in one source only) that MEXT was considering introducing a similar assessment for the end of upper secondary.

⁹ This is comparable to A levels/T levels/BTEC). NCEAs at Level 3 provide the necessary qualifications to meet University entrance requirements for students wishing to go on to HE.

¹⁰ BTEC Level 1 and 2 (equal to GCSEs and can be taken alongside or instead of them (at age 16+). BTEC Level 3 (equal to A-levels and can be taken alongside or instead of them). Vocational qualifications can be taken but there is not a distinct pathway until upper secondary level.

¹¹ Vocational Upper Secondary (age 16+) Education and Training.

¹² Vocational education and training (VET) programmes are available (age 16+). The summative assessment in VET and academic education may not necessarily take place at the same time.

¹³ Can take a vocational exam if go to vocational or sectoral or special school (age 16+).

¹⁴ Further Education and Training (FET) course are available (final two years of Senior Cycle, typically age 16+).

¹⁵ Yes. Preparatory vocational education (lower secondary education, Voorbereidend Middelbaar Beroepsonderwijs – VMBO, ages 12-16), and senior secondary vocational education as well as training (Middelbaar Beroepsonderwijs – MBO, ages 17-18).

¹⁶ Diploma of Vocational Education (DVE), which is a credit-based system with flexible curriculum (age 16+).

¹⁷ Yes. There are three programmes available (age 16+)

¹⁸ Vocational subjects can be studied (age 15+), but it is unclear from the policy sources whether there is a distinct vocational pathway/qualification available.

¹⁹ Vocational qualifications can be taken from NCEA Level 1 to Level 3 (age 15+) and often include structured workplace learning. In New Zealand, students can select from hundreds of individual 'standards' within each of the 6 creative industries to accumulate the number of credits required to pass at each NCEA level.

²⁰ This is the age of compulsory full-time education and is sometimes called the school-leaving age. Most children attend school until this age, but it is legal to get full-time education elsewhere (e.g., home schooling). It is compulsory for young people to be in education *or* training (including apprenticeships) until the age of 18.

²¹ Compulsory education level in Finland is reached when students are 18 years-old or have completed an upper secondary qualification (General Upper Secondary or vocational training qualification)

²² School attendance is compulsory for all children aged from 5 to 16 in the Netherlands. People aged 16-18 have to get the basic qualification (HAVO or VWO) before leaving school.

²³ Divided into subject-specific domains. There is a common component which is core component, a profile component, and a free choice component. The core subjects ensure that students have a broad and general education, while the optional subjects allow them to personalize their curriculum and prepare for their future career.

²⁴ However, there is some evidence to suggest this may vary at the school stage. General subjects are: maths, reading, arts, social studies, history, geography, science, music and PE.

²⁵ Students must complete a total of 30 credits to receive the diploma, of which 18 are compulsory. Students must also complete 40 hours of community service.

²⁶ However, some schools require students to study English or another language-rich subject to meet university entrance literacy requirements.

²⁷ Other assessments may not be standardised.

²⁸ Some is school-based.

²⁹ It is partially standardised as it involves both national and school-based examinations. The national examination is standardised at the end of each level of VMBO, while the teacher assessment and coursework are more varied and depend on the school's own system or external parties' requirements.

³⁰ The NCEA is a national program, but it is not 'standardized' in the sense of having a single set of exams that all students take at the same time. Instead, it is flexible and allows for different pathways depending on the student's perceived strengths.

³¹ Internal assessment is not standardised. External (Matura exam) is standardised.

³² Standardisation is carried out by the Leaving Certificate Examinations Board.

³³ Not standardised at the national level, but rather depends on the type and level of education. Students can choose their own learning profiles/pathways.

³⁴ This is because there is no national summative assessment system at this level.

³⁵ This is because there is no national summative assessment system at this level

³⁶ Between 2022 and 2023. In 2024, full pilots of all TMOA NCEA Level 2 Wāhanga Ako and all NZC NCEA Level 2 subjects will be conducted ahead of the implementation of all NCEA Level 2 Wāhanga Ako and NZC subjects in 2026.

³⁷ No, GCSEs and A levels are not directly used to improve teacher training and/or curriculum, but they may have some impact on them through their role as a measure of student achievement and a reflection of national standards.

³⁸ Used to provide feedback to teachers and students on the effectiveness of the learning process. They are used to support school improvement through formative assessment and digital tools.

³⁹ Used for improving policy.

⁴⁰ Used for institutional accountability and to help educational institutions improve their programmes and identify areas of weakness.

⁴¹ Used particularly for TSA, to enable teachers to understand students' learning problems and needs so that targeted and appropriate teaching practices can be put in place.

⁴² Used to inform curriculum planning by analysing learning patterns of students.

⁴³ Through the publication and use of school performance tables.

⁴⁴ It is used to compare school performance and quality between schools and regions.

⁴⁵ Not explicitly stated as being used for accountability.

⁴⁶ They are used for both regional and system accountability and used to monitor quality.

⁴⁷ The terminal assessments are used for this purpose. The classroom assessments are not.

⁴⁸ They provide data and evidence on student achievement and inform decision-making at various levels, including schools, school boards, and the Ministry of Education.

⁴⁹ Used to identify problematic policy areas.

⁵⁰ No, GCSEs and A levels are not directly used to improve schools/regions and aid allocation of resources, but they may have some impact on them through their role as a measure of student achievement and a reflection of national standards.

⁵¹ Secondary school assessments are not used to improve schools/regions and aid allocation of resources.

⁵² They are used to support school improvement.

⁵³ Although school performance is only a small aspect of resource allocation.

⁵⁴ To identify weaker performing schools.

⁵⁵ NCEA achievement data contributes to school review, enabling schools to take their students' progress into account (against national expectations) when planning school programmes, goals, and targets.

⁵⁶ Our expert reviewer has advised that final year assessments are used to sort students or schools into different categories or tiers based on their grades in practice; however, this was not mentioned in the sources as a stated purpose.

⁵⁷ Lower level is used to assess eligibility for upper or advanced assessments.

⁵⁸ The grades are used for sorting students into different levels or 'streams' based on their academic performance.

⁵⁹ NAAAs at age 15 are not used to determine student progression or qualifications

⁶⁰ From the evidence we gathered, it is not clear in what way assessment is being used to improve progression nor how this relates to grouping/sorting.

⁶¹ Our expert reviewer has advised that final year assessments are used to sort students or schools into different categories or tiers based on their grades in practice; however, this was not mentioned in the sources as a stated purpose.

⁶² They are used to establish and demonstrate students' knowledge and skills.

⁶³ Sorting for progression is a particularly important function for the HKDSE.

⁶⁴ The grades are used for sorting students into different levels or 'streams' based on their academic performance.

⁶⁵ No, secondary school assessments are not published in league tables or official statistics.

⁶⁶ Mixed. The results of internal school exams are not published. The results of summative assessments (such as the Matura Exam) are publicly available.

⁶⁷ Yes at the regional, but not at the institution level. TSA results are only published at a territory-wide level. Results of the HKDSE are released at territory-wide level but not by individual school.

⁶⁸ Yes at the regional, but not necessarily at the institution level. For the NAAA, MEXT publishes national averages and mean subject scores for each of the 7 regions of Japan, and ranks the 47 prefectures accordingly. Education policymakers in all 1718 municipalities have the discretion to publish the results of individual school-mean test scores under their jurisdiction or not.

⁶⁹ NZQA publishes annual secondary school statistics online at a national and regional level (there are 10 regions in NZ). The published data also includes results for literacy and numeracy requirements, NCEA certificate endorsements and course endorsements. Annual school league tables are also published by NZQA.

⁷⁰ Responsible for specifying the national core curriculum.

⁷¹ In collaboration with stakeholders which include regions, schools, parents, and pupils.

⁷² Sets the objectives and framework.

⁷³ Is responsible for setting education policy.

⁷⁴ Establishes the qualification standards, regulates them, and sets the rules for how exams are conducted. Underpinning this is a competitive system of multiple exam boards.

⁷⁵ Is accountable for the progress and implementation of the national core curriculum, including assessment guidance. The Matriculation Examination Board (MEB) is responsible for the matriculation exam.

⁷⁶ Write the national curriculum framework. The Ministry of Education and Research (MER) oversees developing and carrying out national assessment policies in collaboration with the Estonian Qualifications Authority (EQA) (Kutsekoda).

⁷⁷ Develops the curriculum and guidelines for implementation. The National Qualifications Framework (NQF) is used to ensure consistency and coherence in the assessment of learning outcomes across the education system.

⁷⁸ The National Council for Curriculum and Assessment (NCCA) develops the curriculum. The State Examinations Commission (SEC) is responsible for the development, assessment, accreditation and certification of the second-level examinations of the Irish state: the Junior Certificate and the Leaving Certificate. The SEC also has a role in standard setting and quality assurance.

⁷⁹ They work closely with other stakeholders, such as the Inspectorate of Education, the National Institute for Curriculum Development (SLO), the College for Examinations (CvE), and the Education Council, to monitor and evaluate the quality.

⁸⁰ Is responsible for administering the TSA and HKDSE assessments, based on curriculum advised by the Curriculum Development Council (CDC) for TSA and HKEAA for HKDSE core subjects.

⁸¹ Ontario Ministry of Education writes it. The Education Quality and Accountability Office (EQAO), is responsible for evaluating the quality and effectiveness of elementary and secondary school education.

⁸² No, assessments are determined by the DfE, Ofqual, and exam boards.

⁸³ National guidance must be followed, but local authorities, schools, and students can have some choice/influence. Local authorities have some autonomy in managing schools but must follow national guidance. Schools have some flexibility to adapt the curriculum to their context and to meet student needs (e.g. deciding on methods and forms of assessments). Pupils can influence the assessment by choosing some optional subjects or courses according to their interests and abilities.

⁸⁴ For example, the Primary Curriculum Framework emphasises the partnership approach and recognises local diversity.

⁸⁵ For example, schools have autonomy to tailor their education to the needs and interests of their students and communities.

⁸⁶ Local implementation of the assessment can involve locally-designed SBA.

⁸⁷ Including the Finnish National Agency for Education (EDUFI), universities, and teacher training institutions.

⁸⁸ Results of national assessments are used to monitor educational performance of students, guide policy, and improve the curriculum. Results are compared across different types of schools to evaluate the effectiveness of educational policies and practices.

⁸⁹ Through external quality assurance mechanisms, such as the evaluation and accreditation of educational institutions, the evaluation of the education system, and the monitoring and evaluation of educational outcomes.

⁹⁰ Individual schools are encouraged to self-evaluate. National processes such as PISA. Assessment by various internal stakeholder groups.

⁹¹ Combination of external and internal processes, including monitoring by Ministry of Education, Culture and Science, and self-evaluation from schools.

⁹² Through education authorities, continuous improvement processes, collaboration and partnerships, research and evidence-based practices, and accountability measures.

⁹³ Questionnaires are completed by students, parents and schools to provide a view of the relationship between student performance, learning environments, student lifestyles and teaching practices. The Central Council for Education (an advisory body to MEXT) conducts research and produces reports that inform policy decisions.

⁹⁴ NZQA reviews and quality assures the NCEA through a formal process known as the Managing National Assessment Review (MNA) which takes place at least once every 4 years. NZQA follows a standard consultation process for any changes to the NZQF, this process includes consultations with a range of educational professionals and stakeholders including those from the secondary schooling sector, university sector, Institutes of Technology and Polytechnics (ITP) sector and Private Training Establishment (PTE) sector.

⁹⁵ The Polish Accreditation Committee and the General Council for Higher Education are responsible for the external evaluation and accreditation of institutions. The Ministry of Education and Science evaluates the education system, and the National Centre for Educational Development monitors and evaluates educational outcomes. The National Examination Board (Centralna Komisja Egzaminacyjna, CKE) employs a thorough and rigorous approach for evaluating the summative assessment system. They regularly conduct reviews, evaluations, and audits to ensure the quality and fairness of the assessment system.

⁹⁶ There are a variety of stakeholders involved in evaluation. The Inspectorate, a division of the Department of Education, is responsible for evaluating the quality of education provision in schools and centres for education.

⁹⁷ Mostly the Central Council for Education. The Japan Language Testing Association (JLTA) published the Code of Good Testing Practice and in 2007, the Japan Association for Research on Testing (now called the Japan Institute for Educational Measurement) published test standards. However, they are not directly involved in government discussions or decision making on national assessment of academic ability.

⁹⁸ Concerns about the rigour and comparability of qualifications, the need to better prepare students for the demands of higher education and employment, and the desire to raise standards and improve educational outcomes. Emphasis on minimising the quantity of coursework and other forms of internal evaluation in favour of end-of-course tests to increase the reliability and validity of results and limit the possibility of cheating or gaming the system. Desire to improve the use of technology in assessment to make tests more accessible to students with disabilities and lessen the administrative burden on instructors and examiners.

⁹⁹ A growing concern over the ‘increasing inequalities’ in educational outcomes, leading to a push for more basic assessment practices that better consider students’ individual needs and strengths. A greater emphasis on the development of ‘21st century skills’, such as problem-solving, critical thinking. Increased attention to the use of technology in assessment, including the use of digital tools and platforms for assessment purposes.

¹⁰⁰ Increasing recognition that learning outcomes are important and must be monitored through assessment. Estonia’s assessment policies and changes are motivated by a desire to foster the development of core competencies, such as digital and civic abilities, which are regarded as critical for success in the 21st century.

A national commitment to providing a high-quality equitable education for all students. Strong emphasis on formative feedback to inform students and teachers.

¹⁰¹ Increasing recognition that learning outcomes are important and must be monitored through assessment. Estonia’s assessment policies and changes are motivated by a desire to foster the development of core competencies, such as digital and civic abilities, which are regarded as critical for success in the 21st century.

A national commitment to providing a high-quality equitable education for all students. Strong emphasis on formative feedback to inform students and teachers.

¹⁰² The need to align assessment with educational objectives, to promote student-centred learning, and improve educational justice and inclusivity. The importance of balancing summative and formative assessment.

¹⁰³ The desire to improve the quality and relevance of education to meet the changing needs of society and the labour market. to address issues of equity and inclusion, particularly for students from disadvantaged backgrounds. Greater emphasis on student-centred learning and personalised approaches to education, which has influenced the way assessments are designed and used.

¹⁰⁴ Modernising the system by increasing pathways and accessibility through tuition-free vocational options. Widespread concerns amongst students, parents and educational professionals about the intense pressure that students.

¹⁰⁵ The Education Quality and Accountability Office (EQAO) driving transparency, fairness, and supporting students from diverse backgrounds. Also, data-driven assessment making.

¹⁰⁶ Stronger emphasis on summative than formative assessment. A credit system was introduced by MEXT in the 1990s due to concerns about equity and access within the post-compulsory education system. Students who graduate (about 98% of all students) are awarded the Senior High School Graduation Certificate (sotsugyo shomeisho) and are eligible to sit for university entrance examinations.

¹⁰⁷ Emerging concerns around educational inequality and early school leaving, and to ensure that the summative assessment system caters for the diverse needs of students. Some of the key reforms have been to increase the number of vocational ‘standards’ and pathways on offer, greater range and flexibility in subject/course choice and in forms of assessment/credit accumulation and the introduction of curriculum content and qualifications specifically to meet the needs, skills and aspirations of Māori and Pasifika students who represent more than one-third of the student population.

¹⁰⁸ COVID-19 pandemic necessitated a balance between assessment expectations and student flexibility and support.

¹⁰⁹ Finland’s strong focus on equity and social justice have immensely influenced the development of assessment policies and practices in the country.

¹¹⁰ Creating a highly skilled and competitive workforce.

¹¹¹ Globally recognised educational standards. As a member of the European Union, Poland is subject to a range of international agreements and obligations that require it to maintain high standards of education and to ensure that its qualifications are recognised throughout Europe and beyond.

¹¹² Hong Kong's history as both a Chinese and British territory (now a Special Administrative Region of China) means that its educational system and schools are a hybrid of British and Chinese traditions. Recent changes have moved away from a UK system towards a US and Chinese style system. Concerns about low PISA scores on psychosocial factors compared to high academic performance.

¹¹³ Tests in Japan are seen as valuable for the development of student's self-discipline and determination, influenced by the context of Japan's Confucian heritage influence. Juku – after-school tutoring schools (also known as cram schools) - are attended by over half of all Japanese students. This has raised concerns about equity as attending juku is expensive; the Japanese government has tried several policy responses to limit attendance at juku but has not been successful in decreasing their influence. Covid pandemic accelerated plans to integrate technology into classrooms. The most recent 'Course of Study guidance' (2020-22) aims to nurture student's cross curricular competencies (e.g., problem solving), partly in response to international criticism by academic experts and education professionals that independent/critical thinking and problem-solving have not been fostered in Japan's teaching approach or outcome-based assessment.