RESEARCH AND ANALYSIS

A review of approaches to assessing writing at the end of primary education

Drawing upon historical and international practices

ofqual

Author

This report was written by Benjamin M. P. Cuff, from Ofqual's Strategy Risk and Research directorate.

Acknowledgements

I would like to thank all those who provided feedback on earlier drafts of this report, in particular Tina Isaacs from Ofqual's Standards Advisory Group, Rebecca Clarkson from the University of Exeter, members of the Standards and Testing Agency (STA) and members of Ofqual's Research Advisory Group. For helping to quality assure the international review, I would like to thank representatives from the Education Quality and Accountability Office (EQAO; Ontario, Canada), the US Department of Education, the New Zealand Ministry of Education, the Singapore Examinations and Assessment Board (SEAB), the Hong Kong Examinations and Assessment Authority (HKEAA), the Scottish Government Learning Directorate, and the Australian Curriculum, Assessment and Reporting Authority (A CARA).

Contents

Executive summary

Writing is a form of communication inherent to daily life, education, and employment. It therefore constitutes one of the main objectives of early education systems around the world, and a number of different approaches to assessing writing can be taken. The purpose of this paper is to review historical approaches taken in England and those currently being taken abroad, to allow for a consideration of the various advantages and disadvantages of different approaches. In particular, approaches are evaluated in relation to definitions of writing and purposes of assessment. The main aim of this paper is to provide a useful resource to researchers and stakeholders on the different approaches that may be taken. While each assessment must be designed to meet its particular purposes in its own specific context, and there should be no assumption that what works in one jurisdiction or context will work in another, there may be useful learning points both from historical debates and international practices. The report is divided into 5 main parts, as outlined below.

Part 1 sets out how writing might be conceptualised, which is necessary to inform thinking on how different aspects of writing might be covered in an assessment. It considers how writing tends to be more formal than spoken language, is usually produced in the absence of immediate audience feedback, meaning that the writer must anticipate potential issues with audience understanding and/or engagement. It also discusses how writing is usually defined in terms of both technical (eg grammar, punctuation) and compositional skills (eg content and style). While there are several different aspects of writing, assessments need not necessarily focus on all of them. It will largely depend on the purpose(s) of the assessment being designed.

Part 2 reviews the history of writing assessment under the National Curriculum in England, focussing on Key Stage 2 (KS2) assessments taken at the end of the primary stage. In particular, this review shows that a number of different approaches have been taken. For example, KS2 writing was initially assessed both by external tests and by teacher assessment in 1995-2012, becoming solely teacher assessed from 2013 onwards (with a separate grammar, punctuation, and spelling [GPS] test). The main writing assessment focussed on the production of extended responses throughout this time, with the GPS test including short items focused on specific technical skills). A greater number of changes were made in the approaches taken to marking/judging. In summary, it was originally intended that writing should be judged according to several ‘statements of attainment’. However, this was felt to be too burdensome for the first Key Stage 1 and 3 assessments (1991-1994; which pre-dated KS2 assessments), encouraging a ‘tick-list’ approach. As such, a best-fit model, using more holistic descriptors, was therefore adopted for KS2 assessments in 1995-2012. Concerns were then raised that this approach was too flexible, and did not ensure full coverage of the curriculum as intended. A secure-fit (mastery) model was therefore adopted in 2016, which reflected a move back towards the old statements of attainment (each standard was described by a number of statements, all of which needed to be met to achieve that standard). Similar to earlier debates, however, the inflexible nature of this approach again became a concern, leading to the introduction of some flexibility in 2018 (for pupils otherwise securely working at a particular standard but where a particular weakness would prevent an accurate outcome being given under a stricter secure-fit model).

Part 3 reviews 15 international assessments which are described by each jurisdiction as being an assessment of 'writing'. These are all large-scale primarily summative assessments of writing at the end of primary education. Findings demonstrate that a variety of different approaches are taken internationally, for a variety of different purposes (eg to provide information on pupils, schools, or jurisdictions) and in both low-stakes and high-stakes contexts. For the purpose of this paper, 'high-stakes' assessments are defined as those which are used to make pupil progression decisions or contribute towards school accountability measures; 'low-stakes' assessments are defined as those not used for either of these purposes. Most jurisdictions use external tests (some paper-based, some computer-based), and 2 (England and the Caribbean) use teacher assessed portfolios. Most assess writing via extended responses (ie one or more paragraphs in length), but some require a number of short responses from pupils (single words or single sentences) or use multiple-choice type items. Some assessments focus on specific types of writing (eg narrative or informative writing), whereas others do not. Some require pupils to produce a greater amount of writing for assessment than others (eg whole portfolios versus short responses). Finally, differences in the approach to marking/grading/judging were observed, ranging between points-based, secure-fit, or best-fit approaches.

While not identified in the international assessments that were reviewed, Part 4 considers comparative judgement methods (where multiple rank-order judgements are combined via a statistical technique to produce an overall scale of proficiency), and automated (computer) marking of extended responses as notable innovations in the assessment of writing.

Finally, Part 5 draws the preceding sections together to discuss the various advantages and disadvantages of different approaches, focussing on construct (ie how writing is conceptualised for assessment) and purpose (eg the intended uses of assessment outcomes). For example, whether writing is assessed as a complete construct or the focus is on specific skills within writing (eg grammar) has various implications for assessment design, in particular for the mode and type of assessment and the approach to marking/grading/judging. The desired breadth and depth of coverage of different genres in writing will have further implications for the setting of tasks. The intended uses of assessments also impact upon what information outcomes need to provide relating to pupils, schools, and/or jurisdictions. Issues relating to reliability and validity are of course also important considerations, such as potential trade-offs between authenticity and levels of standardisation and control over the assessment, and when considering what the preferred approach to marking/grading might be.

The implications associated with each of these decisions depend to a large extent on individual contexts, aims, and constraints (financial, policy, etc.). This paper does not seek to conclude which approach might be 'best' in assessing writing at the end of primary education. Rather, discussions presented within this report aim to consider how such decisions are made in light of individual contexts.

Introduction

The ability to communicate ideas in writing is one of the 3 key methods of communication, with the others being verbal and non-verbal/behavioural. The ability to write well has particular importance throughout many areas of education and employment. It should come as no surprise, therefore, that writing constitutes one of the main objectives of primary/elementary education systems both in England and abroad. So as to monitor progress and proficiency in writing, many countries include writing in their national assessment programmes. As with any assessment carrying this level of importance, reliable and valid measurement is highly desired.

The focus of this paper is on large-scale (national/state/provincial) primarily summative assessments of writing at the end of primary/elementary education. Those interested in writing assessments developed for research purposes (ie not nationally implemented assessments), relating to both summative and formative assessments, for any age group, are referred to McGrane, Chan, Boggs, Stiff, and Hopfenbeck (2018). However, it is worth noting that many of the same discussions presented within the current paper may also apply to other contexts[1].

There are several different ways to approach the assessment of writing at the end of primary education. The purpose of this report is to present and discuss these different approaches. This report does not attempt to arrive at any conclusions as to which assessment method is 'best', because such conclusions would largely depend upon the purpose and uses of particular assessments within individual contexts. Rather, the aim of this report is to provide a useful resource, facilitating considerations of the various issues at hand in relation to specific contexts, and to discuss the implications those issues may have on assessment design.

In meeting these aims, this paper comprises 3 discussions. The first discussion focusses on a consideration of how 'writing' can be defined, to better understand what might be covered in an assessment. The second discussion presents a history of writing assessments under the National Curriculum in England (1988 to present day), to review and learn from past debates. The third discussion focusses upon current international practices, to again consider what different approaches might be taken. These 3 discussions will then be drawn together in Section 5 of this report, in which the various advantages and disadvantages of different approaches will be discussed in light of how writing might be conceptualised, and the potential intended uses of assessment outcomes.

[1] For example, various writing assessments exist which are targeted at adult populations, mainly those for whom English is a second language, such as:

- TOEFL: https://www.ets.org/toefl
- B2 First (previously known as FCE): http://www.cambridgeenglish.org/exams-and-tests/first/
- IELTS: https://www.ielts.org/

1 Definitions of writing

As the following discussions demonstrate, there are different aspects of writing, each of which may or may not be included in assessment objectives. In this section, we discuss the main aspects of writing as set out in research literature; we do not set out to provide a single definition of 'writing'. Discussions in this section will be drawn upon in later sections of this report.

One consideration is how the particular features of writing differ from other forms of communication. For example, the audience (in this case, the reader) is not usually present at the time writing is produced, whereas in the cases of verbal and non-verbal/behavioural communication the audience is usually present, and can usually give some immediate feedback as to whether or not the message has been understood. This means that writing must usually be constructed without any immediate feedback. Weigle (2002) argues that proficient writers will therefore be able to shape their message appropriately, expressing an awareness of the audience in terms of their likely pre-existing understanding of a topic, and the content that is likely to be the most persuasive. She also argues that writing in a tone of voice appropriate for the audience might also be important in engaging the reader. For example, in some situations a formal tone of voice can add authority to the message, in others it might make it appear inaccessible. When a writer fails to address these elements, the reader may misinterpret, or disregard the message being communicated.

Flower (1979) described the above in a slightly different way, noting the difference between 'reader-based prose' and 'writer-based prose'. Reader-based prose would be more indicative of greater proficiency in writing, in which writers are not only able to express ideas, but are able to transform ideas to address the needs of the reader. Writer-based prose, on the other hand, serves only to express the writer's own thoughts, which may or may not be easily interpretable by the reader.

Writing as a social device is often also defined by social convention (eg see Moffett, 1983, Chapter 5; Weigle, 2002, Chapter 2). For example, such conventions usually dictate that written language, especially when used in education and employment settings, tends to be more formal than spoken language. Due to this relative formality, technical accuracy can be considered more important and more highly valued in written language than in spoken language. Appropriate use of grammar, punctuation and spelling is therefore often valued. However, careful use of creativity can also be important in producing a piece of writing that is engaging and interesting, yet still remains fit for purpose and audience (ie appropriate in relation to the social conventions of the intended purpose/genre and audience).

Odell (1981, p. 103) emphasises the iterative process that writers go through in generating a number of alternatives from which to choose (eg alternative words, sentence structures, semantic devices, etc.), whose definition of competence in writing included "the ability to discover what one wishes to say". He argues that writers in most cases are not able to select from a pre-determined list of options, but the skill lies in being able to generate such options for themselves, deciding upon the most appropriate choices for the task at hand, and going through an iterative process of revision and refinement through writing.

Writing then, is usually defined in terms of both technical (eg grammar, punctuation, and spelling) and compositional skills (eg content and style)[2]. Handwriting and other motor skills could also be considered important, as poor handwriting could impede a reader's ability to extract meaning from a text. Good writers will be able to make appropriates choices, to express these skills in a manner which is fit for purpose and audience. It would be inappropriate here to define what is meant by 'fit for purpose and audience' because this is context-dependent, and will vary by region and social convention (see Weigle, 2002, Chapter 2). The definition of the above elements have been broken down into more detail elsewhere (eg Weigle, 2002, Chapter 2), but this relatively high-level conceptualisation will suffice for the current discussion.

This section has discussed the features of writing as a whole concept. Again, however, assessments do not need to necessarily focus on all elements. One challenge is to decide which aspects of writing to focus upon, according to the purpose of the assessment. For example, some assessments may target general proficiency in writing (thus may focus upon the construct in its entirety), others may focus on more specific, basic skills within writing, such as grammar, punctuation, and spelling. An awareness of the distinction between 'writing' and specific skills within writing can helpfully inform what a particular assessment will measure. This may also to some extent be age dependent, as one could assess writing in different ways for younger and older pupils, possibly focussing on different elements of writing for each.

2 History of National Curriculum writing assessments in England

While the history of primary school testing in England has been documented more thoroughly elsewhere (eg Bew, 2011; Daugherty, 1995; Shorrocks-Taylor, 1999; Whetton, 2009), a summary of relatively recent approaches is provided here to inform current debate. Again, the focus of this paper is on end of primary school assessments, which for England are those at the end of Key Stage 2 (KS2)[3]. Assessment of other key stages is discussed where these can inform discussions relevant to KS2. For clarity, this section is divided into separate time periods, according to the occurrence of major assessment policy changes.

[2] The National Curriculum in England defines writing in similar terms, specifying teaching of "transcription (spelling and handwriting)" and "composition (articulating ideas and structuring them in speech and writing)" (DfE, 2013, p. 15).

[3] In England, education is divided into 5 'key stages' (KS). Broadly, KS1 covers ages 5-7, KS2 covers ages 8-11 (the final 4 years of primary education), KS3 covers ages 12-14, KS4 covers ages 15-16, and KS5 covers ages 17-18. Summative assessments are delivered at the end of each key stage.

Assessments in 1991-2012

The National Curriculum in England was first implemented in 1988, which introduced statutory assessment in primary education. Prior to this[4], there was no national curriculum taught in schools, and no national system of testing at this level. In addition to the introduction of a common teaching and testing framework, the aims of this new curriculum were to raise standards and provide school accountability. The first national KS1 and KS3 assessments were delivered in 1991 and 1993 respectively (Daugherty, 1995), and the first national assessments for KS2 were delivered in 1995 (Bew, 2011). Pilot testing had been carried out in earlier years for each key stage. KS2 assessments covered maths, science, and English (including reading and writing), with each subject being assessed via a combination of both internal teacher assessment and external testing (Shorrocks-Taylor, 1999).

Teacher Assessment

The original intention for the teacher assessment element of writing, set by the Task Group on Assessment and Testing (TGAT, 1988), was for teachers to grade pupils according to a number of attainment targets, each of which consisted of a number of 'statements of attainment' (SOAs) (TGAT, 1988). Indeed, in the first KS1 and KS3 teacher assessments of writing, teachers were required to assign pupils into 'levels of attainment', each of which were described by a number of these SOAs. While there were no statutory requirements for teachers to assess against every SOA, this nevertheless became common practice (Dearing, 1994, Appendix 6), and due to the large number of SOAs assessed (over 100 per pupil across English, maths, and science; Whetton, 2009), this proved to be a time-consuming exercise. It also led to fragmented teaching and learning (owing to the very atomised assessment criteria), encouraging a 'tick-list' approach to assessment (Dearing, 1994, para. 7.11). This approach to teacher assessment was therefore changed in 1995, meaning that the first KS2 teacher assessments adopted more holistic, best-fit 'level descriptors'[5], instead of the overly specific SOAs (Hall & Harding, 2002). Teacher assessments were subject to external moderation throughout this time (and beyond).

The original assessment developers had also intended for the internal and external assessments to be equally valued (for all subjects), with the importance of teacher assessment being repeatedly emphasised. For example, the TGAT Report (1988, para. 60) described teacher assessment as being "a fundamental element of the national assessment system", and Daugherty (1995, p. 15) noted that while external tests were to be "at the heart of the assessment process", their purpose was to "supplement" teacher assessment. In practice, however, it seems as though teacher assessment was given secondary importance to the tests. For example, less interest was paid to teacher assessment by policy-makers, less funding was made available (eg for moderation), and the outcomes of external tests often took priority over teacher assessment outcomes (eg for accountability) (Daugherty, 1995; Hall &

[4] The debates of the 1970s and 1980s leading up to the introduction of the National Curriculum have been documented by Daugherty (1995).

[5] Levels-based mark schemes are where pupils are assigned to 1 of a number of different levels of attainment, with each level defined by a written description of the expected standard. Assessors make best-fit judgements to decide which description each candidate's work most closely relates to.

Harding, 2002; Shorrocks-Taylor, 1999, Chapter 8). Daugherty (1995) proposed several reasons for this: 1) because the external tests required greater central control/organisation, thus drew the greater attention for development; 2) because policy-makers had greater interest in 'summative' outcomes, rather than the primarily 'formative' teacher assessments; 3) because of greater trust in outcomes of standardised external tests.

External Testing

For the KS2 external writing tests in 1995-2002, pupils were asked to write 1 extended piece of writing, with 15 minutes allocated for planning, and 45 minutes allocated for writing. Pupils could choose whether to produce 'information writing' (eg writing an informative leaflet in response to a given prompt) or 'story writing' (writing a story in response to a given prompt) (SCAA, 1997b). Responses were externally assessed according to 'purpose and organisation' and 'grammar', using best-fit level descriptors (eg see SCAA, 1997a)[6]. A 'level 6 test' (also called the 'extension test') also existed, which was a more demanding version of the test targeted at higher-ability pupils.

From 2003 until 2012, pupils were asked to produce 2 pieces of extended writing (1 shorter piece, and 1 longer piece)[7], each in response to a given prompt (eg a picture, or some text), and complete a spelling test (QCDA, 2010; Testbase, 2018). For the extended written responses, tasks targeted one of a variety of genres in each year (eg narrative, opinion, persuasive, informative) (Testbase, 2018). Pupils were no longer given a choice of tasks.[8]

Assessments in 2013-2015

The next major set of reforms came about largely in response to the Bew Report (Bew, 2011), which raised a number of concerns about the external tests that were being delivered. For the writing test specifically, Bew (2011) commented that outcomes were too task specific (ie some genres were easier to write about than others, which affected comparability over consecutive years, and may have disadvantaged some pupils), and that the test was not a true reflection of real-world writing (eg the time pressures of the tests did not allow pupils to take as much care over their writing, to review and edit, or demonstrate creativity, as they would in the classroom). Bew also raised concerns about unreliability in the marking of the tests,

[6] Note: past papers could only be found from 1997 onwards. Past papers or mark schemes could not be found for the 1995 or 1996 test series.

[7] The shorter piece of writing was allocated 20 minutes, and the longer piece of writing was allocated 45 minutes including up to 10 minutes for planning (QCDA, 2010). Having 2 pieces of writing showed that pupils could write for different purposes (Colin Watson, personal communication, March 7th, 2019) – prior to 2003, pupils only produced 1 piece of writing in the test.

[8] This was to allow greater control over the genres covered, to better ensure comparability between pupils and greater reliability in marking, to reduce the time spent by pupils in choosing which question to answer, and to introduce an element of unpredictability to reduce the risk of teaching to the test (Sue Horner, personal communication, January 31st, 2019; Colin Watson, personal communication, March 7th, 2019).

which was later given some support by the findings of He, Anwyll, Glanville, and Deavall (2013)[9]. A greater focus on 'essential' technical knowledge and skills (grammar, punctuation, spelling, and vocabulary) was encouraged, but Bew recommended that compositional skills should be given the greater priority. In response to the above concerns, one of the main recommendations of this report was that 'writing composition' (ie the more creative/extended aspects of writing) should be assessed only via internal teacher assessment, as this would allow for a broader range of genres to be covered than was possible in the test, and would remove detrimental time constraints. For the more technical elements of writing (grammar, punctuation, and spelling), it was recommended that externally marked tests should be retained. Bew (2011) argued that it is easier to mark these aspects of writing as being 'right or wrong', whereas compositional elements tend to be much more subjective.

In 2013, the recommendations of the Bew Report (2011) were largely implemented. External tests were now only taken in reading and maths, along with the newly created grammar, punctuation and spelling (GPS) test (there were no external tests for writing as a whole concept). Level 6 tests were reintroduced for these subjects in 2013, again to challenge and recognise higher ability pupils (Bew, 2011; Testbase, 2018). Teacher assessments still followed a 'best fit' approach, in which pupils were assigned to 1 of 6 'levels'. While an external test now existed to cover grammar, punctuation and spelling, these elements were still also included in the teacher assessment (in addition to compositional skills).

Assessments in 2016-2018

In 2013, it was also announced that the National Primary Curriculum would be revised for first teaching in September 2014 (see DfE, 2013), and first assessment in 2016. Similar to the 1988 reforms, these changes aimed to encourage higher standards and support school accountability (Gove, 2013, 2014), and to ensure that "all children have the opportunity to acquire a core of essential knowledge in key subjects" (Gove, 2013). In response to concerns that the flexible nature of the best-fit assessment criteria contributed to narrowing teaching and learning (because there was no strict requirement to focus on the full breadth of assessment criteria), new ('interim') teacher assessment frameworks were put into place for the 2016 assessments (see STA, 2015). The main change in the approach to assessment was the introduction of 'secure-fit', rather than best-fit, judgements. Similar in nature to the 'statements of attainment' adopted in 1988, this involved the introduction of a number of specific 'pupil-can' statements. In order to achieve each standard[10], pupils needed to demonstrate all of the statements listed within that standard (and the preceding standards), meaning that assessment decisions were deliberately designed to be less flexible (ie more secure) than under the best-fit model. Writing as

[9] Large variation was reported between examiners marking the same 'benchmark' scripts (used to monitor marking consistency), and large differences between examiners' marks and the definitive marks for those benchmark scripts were found.

[10] These were: 'Working towards the expected standard', 'working at the expected standard', or 'working at greater depth within the expected standard'.

a whole subject was still assessed only via teacher assessments, alongside the separate external grammar, punctuation, and spelling test.

Similar to the concerns raised about the 1988 statements of attainment, stakeholders began to express concerns that this new approach was too rigid, and had become a 'tick-box' exercise, increasing workload for teachers (eg see National Education Union, 2016; Schools Week, 2017; TES, 2016). In particular, it was felt that this approach created issues of fairness for pupils with particular weaknesses (eg poor spelling – see House of Commons Education Committee, 2017).

The teacher assessment framework for writing was therefore revised for the 2018 series (see STA, 2017b), giving teachers more flexibility to use their professional judgement for pupils with a particular weakness: where a pupil is otherwise securely working at a particular level, but a particular weakness would (for a good reason) prevent an accurate outcome being given under a stricter secure-fit model, then that standard can still be awarded (STA, 2017b). Some similarities can be seen here with some early thinking in 1989, where an "n minus one" rule was considered by one of the developers of the first KS3 assessments: where a pupil failed to demonstrate just 1 statement of attainment for a given level, they could still be awarded it (Daugherty, 1995, p. 49). Writing composition was also given greater emphasis in the 2018 assessments, making requirements somewhat less prescriptive for technical skills (ie grammar, punctuation, and spelling).

The external grammar, punctuation, and spelling test continued to be delivered during this period. However, due to the stretching nature of the new assessments, the more demanding Level 6 tests were discontinued from 2016. More demanding items were instead integrated into the main test (noted in STA, 2017a).

It is worth noting that different methods of assessing writing at this level are currently (ie at the time of writing) being explored. As can be seen in evidence given to the Education Select Committee (House of Commons Education Committee, 2017), some stakeholders are in favour of retaining teacher assessment, whereas others would like to see a different approach, such as a comparative judgement design (further detail on this is given in Section 4.1).

To summarise this section, Figure 1 shows a timeline of the main changes to the assessment of writing at the end of the primary stage that have been discussed.

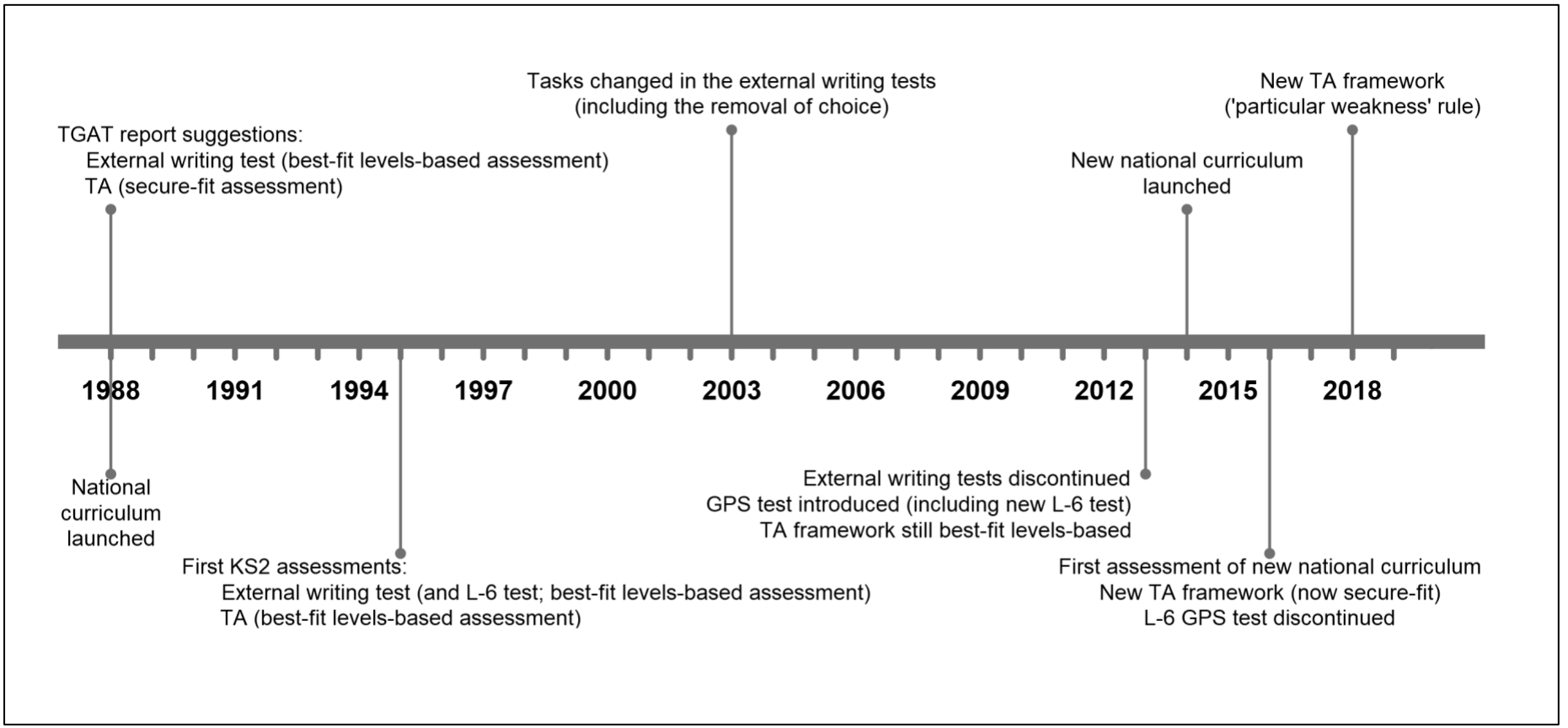

Figure 1. Summary of the main changes to the assessment of writing at the end of primary education in England (1988-2018)

Notes. "TA" = teacher assessment; "L-6" = Level 6 (test); "GPS test" = grammar, punctuation, and spelling test

3 Review of international approaches

3.1 Method

The purpose of this section is to consider what approaches to the assessment of writing are currently being taken internationally. The aim is not to provide a critique on the assessments identified. Rather, they are simply used as a device through which to identify the different ways in which writing can be practically assessed in a large-scale setting. It should not be assumed that arrangements can be transferred across contexts and jurisdictions; this will depend upon numerous factors and so caution should therefore be employed.

Some decisions needed to be made regarding which assessments to include in the review. In keeping with the rest of this paper, the focus was on large-scale (usually national) predominantly summative assessments of primary (elementary) level writing. Focus was not given to any small-scale (ie classroom based) assessments, predominantly formative assessments, or those targeted towards other age groups. Where an assessment targeted multiple year groups, focus was maintained on arrangements relating to the primary school leaving year (eg for England, KS2 assessments, rather than assessments at the other key stages). The review was also only concerned with assessments being explicitly promoted/described as assessments of 'writing'; those explicitly promoted/described as assessments of more specific skills (eg those described as 'grammar/spelling tests') were not included. While related, these are not assessments of 'writing' by intention/design, so fall outside of the scope of this paper. This means that the focus for England here is on the KS2 writing teacher assessment, and not the external grammar, punctuation and spelling test.

There was no provision for the translation of foreign languages, so jurisdictions were identified where English was used for official documents. This included jurisdictions where English is the first language, or where English is not the first language, but is an official language, and therefore used for official purposes. For example, the first language spoken in India is Hindi, but English is used for many official documents. Both England and Scotland were included in the review for the UK. To make the review more manageable, the list was further reduced via the exclusion of any jurisdictions that had a population of less than 1 million (in 2016, according to The World Bank, 2017).

An initial review was conducted on identified jurisdictions (for a full list, see Cuff, 2018, on which the current methodology is based). An online search engine was used to source information on writing assessments, with the aforementioned inclusion/exclusion criteria. Those that did not appear to deliver any writing assessments meeting these criteria were excluded, either because they explicitly only tested specific skills within writing or because no information could be found to suggest the presence of any assessments related to writing. For Sierra Leone, a writing assessment was identified (in the National Primary School Examination), but no further information beyond that could be found for review, and so this jurisdiction was excluded.

The final list of inclusions comprised of 15 identified assessments from 13 jurisdictions (3 were identified in the USA). In Canada and the USA, assessment

practices differ across states/provinces. For these, the largest state/province by population was reviewed; this was to make the review more manageable. In the USA, both a national and 2 state (California) assessments were identified. Some jurisdictions subscribe to multi-national organisations for assessment purposes, such as the Caribbean Examinations Council (CXC – henceforth simply referred to as 'Caribbean'), which offers a writing assessment to member states. Any use of the word 'jurisdictions' in this report should also be taken to include this organisation.

The final list of sampled jurisdictions/assessments included:

- Australia – National Assessment Program: Literacy and Numeracy (NAPLAN)
- Canada (Ontario) – Assessment of Reading, Writing and Mathematics: Junior Division (also known as the Junior Division Assessment; JDA)
- Caribbean – Caribbean Primary Exit Assessment (CPEA)
- England – National Curriculum Assessments: Key Stage 2 (KS2)
- Hong Kong – Territory-wide System Assessment (TSA)
- New Zealand – Assessment Tools for Teaching and Learning (e-asTTle)
- Pakistan – National Achievement Test (NAT)
- Philippines – National Achievement Test (NAT)
- Scotland – Scotland National Standardised Assessments (SNSA)
- Singapore – Primary School Leaving Examination (PSLE)
- Trinidad and Tobago – Secondary Entrance Assessment (SEA)
- Uganda – Primary Leaving Examinations (PLE)
- USA (California) – California Assessment of Student Performance and Progress (CAASPP)
- USA (California) – English Language Proficiency Assessments for California (ELPAC)
- USA (National) – National Assessment of Educational Progress (NAEP)

For each of the above assessments, literature was sought with a number of specific questions in mind. These questions were:

1. What is the main method of assessing writing?
2. What are the main intended uses of the outcomes of the assessment?
3. What are the stakes of the assessment?
4. What specific skills within writing does the assessment aim to cover?
5. How is the assessment marked/graded?

Efforts were made in all cases to glean information from official sources (eg government or exam board websites/reports). However, this was not always possible, and so some media/academic sources were also used where necessary. After sufficient information for each assessment had been found, or at least an exhaustive search had been made, information was organised into a number of tables. These can be found in the appendix and are summarised in the sub-sections to follow. The relevant sections of the tables were sent to the responsible organisation for each of the international assessments, who were given the opportunity to check the statements made within these tables, and to fill in any gaps in information. We received replies from 7 jurisdictions. For the remaining 7

assessments (England was excluded), the documents found online had to be relied upon as representations of how these assessments should be delivered in practice.

3.2 Findings

3.2.1 Purposes and uses of assessments

Firstly, it is worth considering what the intended uses of the assessments are, as these should determine what the assessment outcomes should say about pupils' proficiencies in writing. This in turn should determine what the assessment itself should look like. Readers are reminded that the focus here is on large-scale assessments, which are typically predominantly summative in nature. Predominantly formative assessments will have different purposes and uses, and are not reviewed here.

In general, stated intended uses of the sampled assessments typically fall under 1 or more of 3 types (most fall under some combination; see Table 2 in the appendix):

1. To provide information about pupils – several jurisdictions state that their tests intend to monitor individual performance (in relation to the assessment/curriculum standards and/or other pupils across the jurisdiction), and to identify any children who might need further support. In some cases, assessment outcomes are used to make progression decisions (eg in allocating pupils to secondary schools).
2. To provide information about schools – several jurisdictions aim to use assessments as a quality assurance/accountability mechanism for schools, to monitor their performance, and to allow schools to see how well they compare with others. Assessment outcomes can also be used by schools to help inform/improve their teaching programmes, and to monitor teacher performance.
3. To provide information about the overall performance within the jurisdiction – for example, to produce national data on proficiency in writing (perhaps to compare different demographic groups), to help monitor any changes in pupil ability, and to inform policy decisions. Some also describe using outcomes to understand which schools/areas or types of pupils require greater support, so as to know where to allocate funding.

The uses to which assessments are put also affects the stakes[11] for pupils and schools (see Table 2 in the appendix). Some assessments seem to have very low or no stakes – these include the TSA (Hong Kong), e-asTTle (New Zealand), NAT (Pakistan), SNSA (Scotland), ELPAC (California, USA), and the NAEP (USA). For these assessments, outcomes are used by schools or governments to monitor the

[11] 'Stakes' refers to the extent to which outcomes are used for decision making which is likely to have important implications for individuals (pupils/teachers) and/or schools. For the purposes of this paper, only pupil progression decisions or formal school accountability measures are taken into account when discussing stakes. It is very difficult with a documentation review such as this to capture more nuanced issues, such as the amount of pressure/stress felt by pupils, or the degree of internal accountability within schools. Thus there is likely to be some variation in terms of actual experiences of stakes within each category.

performance of their pupils, but individual or school outcomes may not be made publicly available, and have little or no implications in terms of school funding, scrutiny on schools, or formal progression decisions for pupils.

Other assessments seem to have higher stakes, either for pupils, or schools, or both – these include the NAPLAN (Australia), JDA (Ontario, Canada), CPEA (Caribbean), KS2 (England), NAT (Philippines), PLSE (Singapore), SEA (Trinidad and Tobago), PLE (Uganda), and the CAASPP (California, USA). For these assessments, outcomes may be either explicitly used for pupil progression decisions, or school accountability, or both. In some jurisdictions, pupils failing to achieve a certain grade may not get their desired school placements. In some jurisdictions, teachers/schools failing to meet expectations may face intervention measures, which might include implications for teachers' pay.

3.2.2 Mode and type of assessment

Table 1 in the appendix provides general information on the sampled assessments, including the mode of delivery (paper-based or computer-based test, or a portfolio) and task type (multiple-choice, short-response, or extended-response).[12] Figure 2 summarises this information. As this shows, there is no single model for high or low-stakes assessment of writing. However, the majority of assessments are external tests (ie standardised tasks which are set outside of schools). These are most commonly paper-based, but several are computer-based/typed. Two out the 15 assessments are not externally assessed, but are portfolios assessed in schools by teachers (KS2 [England] and CPEA [Caribbean]). In contrast to the external tests, where pupils respond to the same questions under controlled conditions, portfolios are collections of writing produced over time (eg throughout the school year), with the types of writing included differing across schools. Both of the portfolio-based assessments are high-stakes, as are 7 out of the 13 external tests (5 of the paper-based tests, and 2 of the computer-based tests).

Of the 7 jurisdictions that responded to our request for further information, 5 provided reasons for why they had chosen to use an external test for their main summative assessment. Reasons included that external tests allow for consistent/standardised measurement of educational targets across all schools, with 1 noting the importance of being able to assess all pupils in the same manner within a high-stakes context. Some noted that an external test is used to complement (ie to provide additional evidence of achievement), rather than replace, ongoing teacher judgements.

Across all modes of delivery, pupils are most commonly asked to produce 1 or more extended responses (defined here as writing of at least 1 paragraph in length). The most common approach is to provide some sort of prompt (eg the start of a story, some facts, or an opinion), to which pupils are asked to respond (eg to write a story, to produce a newspaper article, or to discuss an opinion). The next most common type of assessment includes a mixture of task types. For example, the JDA (Ontario, Canada) contains a mixture of extended-response type items (as above) and multiple-choice (eg testing grammar or sentence structure). Just 1 assessment (PLE [Uganda]) only contains short response type items (ie writing a single word or a single sentence, for example to test word choice, spelling, or sentence structure).

[12] Links to example tests/test items are also given in Table 1 where found.

Just 1 assessment is purely multiple-choice (NAT [Philippines]), which is to assess pupils' ability to "identify cohesive devices, identify correct bibliographic entry, [and to] fill out forms" (Benito, 2010, p. 17)[13]. Extended-response, short-response, multiple-choice, and mixed task types were all used in high-stakes contexts.

Figure 2. *Frequencies of the different modes and types of assessment, by stakes*

Note. See Footnote 11 (p.16) for how stakes have been defined within this report.

In all of the assessments reviewed here, writing is assessed as part of a wider assessment suite including other subjects such as reading, maths, science, and/or social studies. Writing is often included as part of a wider 'language' assessment (usually constituting reading and writing). Some assessments focus on single year groups, whereas others include multiple year groups. Most aim to assess every pupil in the relevant year-group(s), either throughout the jurisdiction or throughout those schools subscribing to the assessment. However, the NAT (Pakistan) and the NAEP (USA) (both low-stakes tests) assess only a sample of students and schools[14].

Two of the computer-based tests are 'adaptive', meaning that pupil performance on items earlier on in the test determines the levels of demand of items presented later

[13] Actual examples of test papers could not be found, so further detail cannot be given.

[14] For the NAT (Pakistan), 50% of government schools are sampled, and then 20 pupils are sampled from each school (selected by the Pakistan Ministry of Federal Education and Professional Training, 2016a). For the NAEP (USA), a nationally representative sample of schools is sampled, according to region, ethnic composition, and pupil achievement. Around 10 pupils are then sampled per grade per subject (US Department of Education, personal communication, December 31st, 2018).

in the same test. In other words, if pupils do not perform very well on the first few items, then the test begins to present less demanding questions; higher performing pupils will be presented with more demanding questions. Because these decisions are made automatically by a computer through an algorithm, they apply to assessments containing multiple-choice type items and/or short-response type items within computer-based tests (which therefore tend to focus on technical skills such as grammar, punctuation and spelling). These are the mixed-type, computer-based tests in Figure 2: the SNSA (Scotland; low stakes) and the CAASPP (California, USA; high-stakes). The advantage of this method is that it allows the test to focus on each pupil's level of ability, without presenting too many questions that are too easy or too difficult for them (eg see SNSA, n.d.-b).

3.2.3 Skill coverage

Table 3 in the appendix outlines what specific skills in writing each assessment aims to cover. This is often determined by curricula, and coverage falls under 3 main categories, outlined as follows:

1. Some assessments seem to have a particular focus on writing for specific purposes. These include the NAPLAN (Australia), NAT (Pakistan), SEA (Trinidad and Tobago), ELPAC (California, USA), CAASPP (California, USA), and the NAEP (USA). Each of these defines a specific genre of writing pupils are expected to be able to demonstrate proficiency in, such as narrative, persuasive, or informative writing. Mark schemes, where found[15], are similarly targeted, with the assessment criteria being specifically focussed on the different genres of writing being produced. For some (eg the NAPLAN [Australia] and the SEA [Trinidad and Tobago]), only one genre is assessed each year, with different genres being assessed in different years, whereas for others (eg the ELPAC [California, USA], NAEP [USA]), the same few genres appear to be assessed each year.
2. Other assessments also assess pupils' proficiencies in writing for different purposes, but have a less specific focus on this requirement. These include the JDA (Ontario, Canada), CPEA (Caribbean), KS2 (England), TSA (Hong Kong), e-asTTle (New Zealand), and the PSLE (Singapore). There is often an expectation in these assessments that pupils should be able to write for a range of different purposes, but what that range constitutes is less well defined or restricted than in the above (eg there is no specific requirement to cover certain genres). Similarly, mark schemes for these assessments, where found[16], were not specific to any particular genre of writing. The main focus in these types of assessments, therefore, is on skills in writing more generally, with an expectation that these skills should be demonstrable across a range of (non-specific) contexts.
3. Other assessments seem to focus entirely on demonstrating specific skills (eg grammar, punctuation, spelling), having very little or no focus on writing for particular purposes/audiences. These include the NAT (Philippines), SNSA

[15] Mark schemes for the NAT [Pakistan] could not be found.

[16] Mark schemes for the CPEA (Caribbean), NAT (Philippines), and the PSLE (Singapore) could not be found.

(Scotland), and the PLE (Uganda). Because these assessments are based on multiple-choice type items or single word/sentence answers, they do not appear to address context-specific skills in writing. Note: this does not necessarily mean that other skills are not assessed in any way in these jurisdictions, just that these skills are not covered in their main summative assessments that were reviewed.

3.2.4 Size of the assessments

Table 3 in the appendix also outlines the amount of material that students are expected to produce during the assessment. Of course, this is largely dependent upon what types of items are included and the minimum amount of material needed to allow for sufficiently valid assessments to be made.

Unsurprisingly, those tests that are based upon multiple-choice type items or short-response type items elicit the least amount of material from pupils (ie where pupils only tick boxes or produce a series of single words/sentences: the NAT [Philippines], SNSA [Scotland], and the PLE [Uganda]). Where a greater amount of material is produced, in extended-response type items, differences were observed both in terms of breadth (number of responses) and depth (length of responses). To give some examples, several external tests only ask pupils to produce 1 piece of extended writing with a 40-50 minute time limit (eg the NAPLAN [Australia; high-stakes], TSA [Hong Kong; low-stakes], e-asTTle [New Zealand; low-stakes], and the SEA [Trinidad and Tobago; high-stakes]). The NAEP (USA; low-stakes) and the PSLE (Singapore; high-stakes), however, require pupils to produce 2 pieces of writing with a 60/70 minute time limit, and the ELPAC (California, USA; low-stakes) has 4 items (2 short-response, 2 extended-response) with no official time limit. Those assessed via portfolio (CPEA [Caribbean], KS2 [England]; both high-stakes) likely reflect the greatest amount of material to be produced, both in terms of breadth and depth.

3.2.5 Marking and grading

Table 3 in the appendix also outlines marking and grading practices. Again, this is largely driven by the mode and type of assessment. For example, portfolios in the sampled assessments are marked/graded internally by teachers (KS2 [England], CPEA [Caribbean], both high-stakes and subject to external moderation), whereas paper-based tests are marked externally by examiners (a mixture of high and low stakes). Three computer-based tests use human examiners (the NAPLAN [Australia], which is high-stakes, and the e-asTTle [New Zealand] and the NAEP [USA], which are low-stakes). The SNSA (Scotland, low-stakes), which contains only multiple-choice type items[17] and short-response type items, is auto-marked by a computer. The CAASPP (California, USA; high-stakes) and the JDA (Ontario, Canada; high-stakes) use auto-marking for the multiple-choice items, and external human examiners for the extended-response type items.

[17] Including matching of items and drag and drop selection.

The majority of assessments which include extended response type items (10 out of 11)[18] use some form of best-fit level descriptors for those items, where pupils are assigned to 1 of a number of levels of attainment, based on fairly holistic descriptions of what a pupil is expected to produce ('level descriptors'). Where actual mark schemes were found, the majority took an 'analytical scoring' approach, where separate marks are given for different skills (eg clarity, organisation, etc). The ELPAC (California, USA) and the NAEP (USA) take a 'holistic scoring' approach in their mark schemes, where assessors rely on their overall impression, giving a single score for each piece of writing, rather than separate scores for each individual skill.

A different approach is taken for the KS2 assessment (England). Pupils are not assessed according to level descriptors, but are instead assessed against a secure-fit, mastery-based model. Here, assessors are given a list of detailed criteria, all of which must be met (essentially under a fail/pass judgement) in order to secure a particular standard (although some exemptions can be made for pupils with a particular weakness – see Section 2).

4 Other innovations

Other novel approaches to writing assessment exist but have not yet been adopted in any of the international jurisdictions reviewed in this report. As new approaches and technologies develop they may prove capable of supporting large-scale assessments of writing and so are also worth mentioning in this report.

4.1 Comparative Judgement

One such approach is that of comparative judgement, which has been recommended by some as an alternative to the current KS2 writing assessments in England (eg see House of Commons Education Committee, 2017). In this approach, pupils are not 'marked' in the traditional sense but are rank-ordered via holistic comparisons of 'quality'. The main idea is that it is easier to make relative judgements of quality (eg deciding that one script is better than another) than absolute judgements of quality (eg assigning a numerical score to a script) (derived from Thurstone, 1927). Usually, assessors are shown 2 pieces of writing, and are asked to decide which represents the 'better writing'. Specific criteria are generally not provided, meaning that this method relies upon assessors' pre-existing understanding/beliefs about what good writing looks like. This also means that it is often not possible to know exactly how assessors are differentiating between different levels of performance. After each piece of writing has been subject to multiple comparisons, those comparisons are combined (using a statistical technique known as 'Rasch modelling'[19]) to produce an overall rank-order of pupils in the

[18] The CPEA (Caribbean) is not included in this number – although this contains extended-response type items, information on marking could not be found.

[19] The application of Rasch modelling to comparative judgement was first stated by Andrich (1978). At a very simplified level, the quality of a given script is derived from the number of times it is judged 'better' or 'worse' than other scripts, taking into consideration the estimated quality of those scripts.

cohort (for more information, see Pollitt, 2012a). Once the scale has been produced, cut-off points could be decided upon, from which grades might be assigned.

Some variations on the above have been suggested, mainly to help reduce the number of comparisons that need to be made. For example, instead of making multiple paired comparisons, scripts can be rank ordered in packs of 10 and then that rank order can be converted into multiple sets of paired comparisons (Black & Bramley, 2008), which can then be used for Rasch modelling. Because fewer direct comparisons need to be made, this exercise is less burdensome for judges than traditional approaches. Alternatively, comparative judgements can be used to produce a scale of proficiency in writing, then allowing for the identification of benchmark scripts within that scale. Assessors can then decide which of these calibrated benchmarks each subsequent script is most similar to, meaning each subsequent script only needs to be assessed once, rather than multiple times (for more detail, see Heldsinger & Humphry, 2010, 2013). 'Adaptive comparative judgement' (ACJ) is another alternative, which aims to be more efficient in deriving proficiency scales using a smaller number of comparisons (see Pollitt, 2012a, 2012b). However, caution should be employed with ACJ, as reliability coefficients may be artificially high (ie give an inflated sense of reliability; Bramley, 2015; Bramley & Vitello, 2019).

4.2 Automatic Essay Scoring (AES)

Human marking/judging of extended responses can pose various concerns regarding logistics, ongoing financial cost, and marker reliability. Computer marking via Automatic Essay Scoring (AES) potentially reduces the need for human markers. While auto-marking has already been employed in several jurisdictions for assessing technical skills in writing via multiple-choice and short-response items (eg the JDA [Ontario, Canada], SNSA [Scotland], and the CAASPP [California, USA]), automatic marking of extended responses reflects a greater challenge[20]. This is because extended response type items do not lend themselves to 'right or wrong' answers in the same way as multiple-choice/short response type items do. Nevertheless, some advancements have been made in AES. For example, trials have been conducted for writing tasks in the NAPLAN (Australia) with some apparent success (eg see ACARA, 2015; Lazendic, Justus, & Rabinowitz, 2018). However, these methods largely rely on an analysis of mathematically based textual features (eg vocabulary/sentence length and complexity; Perelman, 2017), and as such, the ability of AES systems to target deeper compositional type skills has been called into question (eg by Perelman, 2017). For example, AES may struggle to recognise skills in creativity, reader-based prose, and persuasiveness. While AES may therefore show some promise, concerns over validity may be too great for some at present.

It is worth noting that AES need not necessarily be used to *replace* human markers but could potentially complement them, by being used as a marker monitoring tool. For example, it could be used to flag human-computer mark discrepancies for further (human) scrutiny (eg as discussed by Whitelock, 2006).

[20] Note that auto-marking of technical aspects of writing is still not infallible – see Perelman (2017).

5 General discussion

As demonstrated throughout the preceding sections, several different approaches can be taken to the summative assessment of writing at the end of primary education. Various approaches have been used in England alone since the introduction of the National Curriculum in 1988. Specifically, KS2 was assessed via both external and teacher assessment between 1995 and 2012, with the former perhaps being given greater precedence than the latter. Teacher assessment then became the main method from 2013 onwards, supplemented with an external grammar, punctuation, and spelling test. Teacher assessments were originally based upon specific 'statements of attainment', in practice taking a secure-fit approach for the first KS1 and KS3 assessments in 1991 and 1993 respectively. However, the first KS2 assessments made use of best-fit judgements based on level descriptors in 1995-2015. This then changed to secure-fit judgements based on specific statements of attainment ('pupil-can' statements) for 2016-2017, and then secure-fit judgements (still based on specific statements) with greater flexibility in 2018. Changes such as these can make maintaining assessment standards more difficult. An awareness of historical debates and changes, including any issues which have surfaced more than once (eg the inflexibility of basing assessments on secure-fit statements), can be helpful to provide longer-term stability in assessment design.

In the international literature, further variety can be observed. Unlike the current preference for teacher assessment in England, the majority of other jurisdictions currently assess writing via an external test (in both high and low-stakes contexts): some paper-based, some computer-based. While the majority use extended-response type items (requiring a response of at least one paragraph in length), some are based upon other item types, such as short-responses (single words/sentences) or multiple-choice. Some assessments focus primarily on writing for specific purposes (eg narrative or informative writing), some have an expectation that pupils should be able to write for a range of purposes (in a less specific manner), and others have very little or no focus on writing for a particular purpose. In some, pupils produce a relatively small amount of material for assessment (eg multiple-choice tests); in others, they produce a relatively large amount (eg portfolios). Most assessments of extended responses adopt a best-fit level descriptors approach (ie where assessment decisions are made according to fairly holistic descriptions of attainment), whereas one (England) uses a secure-fit model (specific 'pupil-can' statements). Finally, variation also exists in the intended uses of assessment outcomes, in terms of providing information on pupils, schools, and/or jurisdictions. Some assessments are used for high-stakes purposes, whereas others are not. While not currently used in any of the reviewed jurisdictions' summative assessments of writing at this level, comparative judgement has been identified as another possible approach, as has automatic essay scoring. Both of these may be worthy of further exploration.

As emphasised in the introduction, the purpose of this paper is not to decide which of these approaches is 'best', as this will depend upon a particular assessment's purpose and skill coverage (ie the assessment construct). The remainder of this section considers these factors in more detail.

5.1 Assessment purpose and construct

The first stage in any assessment design process is to decide upon the purpose of the assessment, including what construct should be measured and how the outcomes of the assessment should be used.

As found in the international review (Section 3), assessments are usually used to provide information on the performance of individuals and/or various aspects of the education system. For example, outcomes can be used to provide information on pupils' progress or attainment, in order to identify those who need further support or to inform progression decisions. They can also be used to provide information on teachers and/or schools as accountability measures, to identify under-performing schools so as to take intervening action, and/or to provide teachers with formative feedback on their teaching practices. Another purpose might be to provide information on a jurisdiction as a whole, to monitor any overall changes in proficiency, to inform policy decisions, and/or to know where to allocate greater funding (ie for certain areas/regions, or certain demographic groups). An assessment may have a number of purposes, which might include any combination of the above. Each intended purpose/usage will have implications for the stakes and design of the assessment, and will need to be compatible with any other purposes and uses.

The extent to which an assessment's purposes can be met will depend upon which approach to assessment is chosen. For example, one of the key aims of the TGAT (1988) for the first national assessments in England was for assessments to have formative benefits on learning, by providing direct information on pupils' proficiency in relation to specific criteria. The intention was for assessments to both feed-back to pupils and teachers about what pupils can do, and where improvements can be made, and feed-forward the same information to the next school (TGAT, 1988, paras. 32–37). Clearly, the choice of assessment method will determine the extent to which outcomes are able to fulfil such intentions, in particular the extent to which outcomes are linked to well-defined assessment criteria. For some assessments, however, such detail might not be necessary. For example, where outcomes are used simply to inform progression decisions, a simple rank order of pupils might suffice.

Another key element informing any assessment design is the definition of the construct to be assessed (ie the skills that should be covered in the assessment objectives). In Section 1 the distinction between 'writing' (ie as a complete concept) and 'specific skills within writing' was discussed. Assessments aiming to focus only on specific skills usually target the more technical elements of writing, such as conventions of grammar, punctuation, and spelling. While such assessments may not cover writing as a complete concept, it may well be decided that technical skills should form the main focus. Assessments targeting writing as a more complete concept are likely to include aspects of compositional type skills among their assessment objectives, such as the ability to write for a particular purpose/audience. For these types of assessment, various other considerations might need to be made, such as what the desired coverage of different genres of writing should be.

Decisions about the purpose and use of an assessment, and the construct being measured, will have various implications for the approach that might be taken. Some modes of assessment and types of items/tasks may be better for meeting certain purposes than others. Some approaches to marking/grading/judging may also be

preferred over others, as different choices here can have different implications for the reliability/validity of outcomes. Such implications should be kept in mind throughout the lifespan of an assessment, not just at the design stage. For example, where the uses of assessment outcomes shift away from original intentions, and/or the stakes of the assessment change, the approach that was originally designed may no longer be a valid way of meeting these new uses.

5.2 Implications for assessment design

5.2.1 Mode of assessment

Most debates about the mode of writing assessment tend to focus on the distinction between internal assessments (eg portfolios) or external tests, and usually concern the trade-off between levels of authenticity (the extent to which tasks reflect real-world writing) and reliability (levels of standardisation and control over the assessment). An appropriate balance may need to be struck in terms of a preference towards more reliable or more authentic/valid assessments, depending on the assessment's purposes, stakes, and how outcomes are used.

Internally assessed portfolios are often considered more authentic than external tests, therefore potentially offering more valid outcomes, because they allow students to produce writing across a range of genres in a similar manner to how they would write outside of an assessment (in part because they tend to have fewer time restrictions). However, there is often little standardisation of tasks, meaning that genre, demand, and levels of teacher assistance will vary between pupils and schools (Koretz, 1998). So, while potentially being more authentic, the use of portfolios may introduce risks to reliable assessment, which may not be desirable in high-stakes contexts. External tests, however, generally present the same tasks to all pupils, and can be marked by a smaller pool of markers who can be more thoroughly trained and standardised, supporting more reliable assessment. However, the time limited nature of external assessment may raise stakeholder concerns over the authenticity of the assessment environment. For example, it may be felt that time restrictions could allow less opportunity for pupils to revise and edit work, or demonstrate creativity (eg see Bew, 2011).

While a distinction between 'reliable tests' versus 'authentic/valid internal assessment' is often assumed to be the case, other factors can affect the extent to which this holds true. For example, while external tests are usually considered to be the more reliable option from a marking reliability point of view, where they sample a small number of tasks, a different selection of tasks for the same pupil may have resulted in quite a different outcome, thus having lower (sampling) reliability (Harlen, 2007). Because portfolios usually contain a greater number of pieces of writing, they may be less susceptible to this issue[21]. This recognition of tests being more reliable than internal assessments may therefore only hold true to the extent that tests are delivered with a sufficient coverage of the construct. Of course, however, the number

[21] A separate issue may exist for portfolios where collections of work are built up over time. As these tend to show progress over time, some judgements may end up being based upon work which does not give an up-to-date demonstration of ability. This is why moderators for KS2 assessments tend to focus on more recent pieces of work (see Cuff, Howard, Mead, & Newton, 2018).

of pieces that can be feasibly assessed will depend upon what pupils can manage within the time available (eg in terms of attention and stamina) the amount of resource available (eg in cost and marking time).

In addition, while outcomes of internally assessed portfolios are often considered to produce more valid outcomes (due to greater authenticity), it is possible that teachers' existing knowledge of their pupils may introduce bias into their summative judgements (eg due to contextual knowledge of a pupil's background and/or past performance) (Cooksey, Freebody, & Wyatt-Smith, 2007). Subjectivity in judgement can have a negative impact on both the reliability and validity of judgements being made at a national level. Teachers may also be more lenient than external examiners when marking/grading (eg Harlen, 2004; McGrane, Chan, Boggs, Stiff, & Hopfenbeck, 2018), which may be a particular issue in high-stakes accountability contexts, where there are incentives to maximise outcomes (House of Commons Education Committee, 2017, paras. 29–30). While moderation can help control outcomes, shortcomings have been identified in the strength of moderation systems as quality assurance processes (Cuff, 2017; Cuff, Howard, Mead, & Newton, 2018). Because external tests are externally marked, where scripts are usually anonymised, there are very little or no opportunities for bias and/or conflicts of interest towards or against particular pupils or schools, thus focus is maintained on validity in relation to national standards. This is ultimately usually what assessment designers intend, and what stakeholders require.

The choice between external tests and internal portfolios might, on the face of it, seem to be a binary one. However, such a distinction need not necessarily always be made. For example, a portfolio with standardised tasks could still maintain authenticity (ie where writing is still produced within the classroom) but have better controls regarding task-setting and marking. Portfolios could also be externally assessed to further increase controls where desired, although that would also increase the cost of marking. For external tests, while some reasonable time restrictions are probably needed, they may not necessarily need to be time limited to the extent that they usually are. Not imposing time restrictions (as is the case for the SNSA [Scotland] and the ELPAC [California, USA]) might help avoid some of the concerns relating to the authenticity of timed tests, such as those raised by Bew (2011). Standardised tests could be internally assessed to bring down the cost of external marking, but this would perhaps have few other advantages, as this would lead to some loss of control over outcomes (internally assessed standardised tests are more commonly used for primarily formative assessments, where such controls are less important).

5.2.2 Item types and prompts

Assessment purpose and construct also has implications for the appropriateness of different item types. For example, where an assessment focusses only on specific technical skills, then multiple-choice (including matching type items etc.) or short answer questions may be sufficient. However, where an assessment focusses on writing in a more complete sense, then extended responses may be better at allowing for a demonstration of higher-level compositional skills.

Nevertheless, this does not mean that an assessment must just use 1 method, as a combination of approaches could be used. Throughout this report, discussions have

tended to treat the different types of assessments in mutually exclusive terms. While this is often the case (eg most international jurisdictions have chosen just 1 method for their summative assessments), different methods can be used in combination to provide a more comprehensive account of proficiency in writing. Some examples include the JDA (Ontario, Canada), and the CAASPP (California, USA), which incorporate a mixture of multiple-choice and extended-response type items. As mentioned in Section 2, taking a combined approach was also something promoted by the task group set up to help design the first National Curriculum Assessments in England (TGAT, 1988), which may help explain why England has used a combination of external testing (the early writing tests, and the current grammar, punctuation, and spelling test) and teacher assessment since the introduction of the National Curriculum in 1988.

The intended skill coverage of an assessment will also have implications for the writing/setting of prompts, where these are used to elicit extended pieces of writing. For example, relatively simplistic, open-ended prompts could be sufficient for the assessment of some technical skills, as the content of the writing produced may be less relevant than its technical features. However, assessment of persuasiveness or narrative writing might require prompts designed to elicit persuasive arguments or story-telling (Weigle, 2002, Chapter 5). Care may need to be taken to avoid overly complicated prompts, however, where one wants to avoid outcomes being affected by pupils' reading abilities (Weigle, 2002, Chapter 5). Where pupils are allowed a choice over different prompts/tasks, care also needs to be taken to ensure that optional routes have comparable demands (Bramley & Crisp, 2019).

5.2.3 Marking/grading/judging

The reliability and validity of an assessment's outcomes is largely influenced by how assessments are marked/graded/judged. This should be derived from the intended purpose of assessment and the construct being assessed. For example, while any assessment method could allow for the separation of the lowest from the highest attaining pupils, different methods offer different degrees of information on pupils/schools/the jurisdiction. Reliable and valid judgement is desired in any assessment, but may be particularly important/desired in high-stakes contexts, as it would be unfair to base any high-stakes decisions for individuals (pupils or teachers) or schools on invalid and/or unreliable outcomes.

Similar to that noted earlier, while different methods tend to be discussed in largely mutually exclusive terms, a combination of methods could be used. For example, Heldsinger and Humphry (2010, p. 14) argued that the best method of establishing validity would be to combine comparative judgements with those based upon mark schemes, to "cross-reference the two sources of information… and to identify anomalous information… in the interests of individual students".

Validity of outcomes in relation to assessment purpose and construct

The extent to which an assessment's outcomes need to reflect the full breadth of the construct may help drive the choice between different methods of marking, grading, or judging. For example, secure-fit models can be used to ensure that assessors evaluate students' writing across the entire breadth of the assessment construct

because assessment objectives are clearly defined, and pupils must demonstrate proficiency in all criteria to achieve each standard. However, this approach can be quite restrictive and time-consuming (see Section 2). In addition, if pupils are meeting (or exceeding) expectations in the majority of areas, but fail to meet expectations in 1 area, they will receive the lower grade (as previously discussed by Cresswell & Houston, 1991). Outcomes from this approach therefore might not always provide a valid reflection of general levels of performance, which might give rise to tensions when outcomes are at odds with teachers' professional judgements (see Section 2). Best-fit approaches are less strict in their focus on every individual criterion, thus can avoid this particular issue, but they do not guarantee such an all-encompassing coverage of assessment objects as secure-fit models.

Comparative judgement approaches produce an overall rank-order of pupils without the need for potentially complicated marking criteria which may not be consistently interpreted. However, while the assessment construct can be defined, this method relies on assessors' holistic understanding of what different levels of performance against that construct look like. In other words, instructions are not given to assessors under this method for how to differentiate between levels of performance in the same way as is done under secure-fit or best-fit models. It is therefore difficult under this approach to know how assessors are making their judgements, and therefore how closely those judgements may or may not reflect the desired depth and breadth of the construct under consideration (eg see van Daal, Lesterhuis, Coertjens, Donche, & De Maeyer, 2019). This could potentially raise concerns over the validity of judgements, and also limit opportunities for feedback. To help with this, assessors could be asked to provide annotated justifications for their decisions, which could be used as part of quality assurance processes. To increase the amount of feedback for pupils and/or schools, annotated exemplars of performance could be produced for various points on the ability scale that this process produces (although this is unlikely to provide the same level of detail on what pupils can do as a clear mark scheme would; Heldsinger & Humphry, 2010, 2013).

The intended skill coverage of the assessment will have further implications for the choice between different methods. For example, where the intention is to focus on technical skills via multiple-choice or short-response type items, a points-based mark scheme may be appropriate (where every mark that can be awarded is clearly defined in terms of what the 'correct' response is). However, different methods may be better suited for the valid assessment of more subjective skills, such as writing for a purpose or audience. Accordingly, most assessments in the international literature have adopted a best-fit levels-based approach. However, while KS2 assessments in England have historically been assessed via best-fit, a secure-fit model was introduced in 2016, because it was felt that the previously adopted best-fit levels-based approach was too flexible (discussed in Section 2). Comparative judgement offers an innovative alternative for judging extended responses. However, while comparative judgement approaches may be effective at assessing writing in a uni-dimensional manner (ie to produce a single outcome score), multi-dimensional assessment (eg to produce separate scores for technical and compositional type skills) is not possible without multiple rounds of judging.

Reliability in marking/grading/judging

In addition to ensuring that the approach to marking/grading/judging allows for valid measurement of the assessment construct, and therefore that outcomes are fit for purpose, it is also important to consider how reliable the outcomes of different approaches are likely to be. Reliability in assessment may again be particularly important when outcomes are used for high-stakes purposes.

Where points based mark schemes are used (eg for multiple-choice or single word answers), there is the potential for highly reliable marking. These types of mark schemes are usually used to assess items which have 'right or wrong' answers, meaning there is often limited room for any misinterpretation of the mark scheme. These kinds of items can also often be automatically marked by a computer, further reducing risks of inconsistency/unreliability. However, as previously noted, these types of mark schemes are often not appropriate for the assessment of deeper compositional type writing skills, limiting their usability in the assessment of writing.

For extended-response type items, the more traditional way of promoting reliability in marking/grading/judging is to put into practice a set of clear assessment criteria, in combination with good training for markers, standardisation and ongoing monitoring and quality assurance. In addition to these factors, the quality of marking will be largely dependent upon the nature of items/tasks included within the assessment, and the item tariff. In general, however, the more clearly assessment criteria are understood, the less scope there is for unreliability. As already noted, standardised tasks make it easier for assessors to evaluate more reliably, because more specific assessment criteria can be produced. It may also easier to train/standardise markers for external assessment compared to internal assessment, where it may not be feasible to train/standardise such a large number of assessors (ie all classroom teachers) to the same degree. Automatic Essay Scoring (AES) can be another potential way of improving the reliability of outcomes. While limitations may preclude its use as a standalone system at present (particularly for judging compositional type skills), AES could still potentially be used as a marker monitoring tool to improve reliability via that process (as discussed in Section 4).

One might argue that a secure-fit model could allow more reliable assessment than best-fit or comparative judgement approaches. So long as assessment criteria are clearly defined, and well understood by assessors (through proper training, etc.), there should theoretically be less room for inconsistency in individual judgements. However, it is very difficult to write criteria that are both detailed and clear, yet still generalise across different pupils and tasks (eg Cresswell & Houston, 1991), and as previously mentioned, any errors or differences of opinion under a secure-fit model can have large consequences on outcomes, thus unreliability. The time consuming nature of secure-fit (see Whetton, 2009) could perhaps increase the risk of errors, as does the fact that a greater number of individual judgements need to be made for each piece of work under this approach compared to others.

Best-fit levels-based approaches offer some advantage in this regard, in that assessment criteria are less restrictive. Of course, while avoiding some of the pitfalls associated with secure-fit, this additional flexibility does introduce other potential risks of unreliability. Concerns over unreliability in marking was one of the reasons why the external writing test, which used a best-fit levels-based approach, was no longer used in England from 2013 (see Section 2). There are different types of best-fit mark schemes (eg see Ahmed & Pollitt, 2011), which may offer varying levels of

reliability. As seen in Section 3, the majority of international assessments seem to adopt 'analytical' mark schemes, where assessors decide upon a separate score for a number of levels-based criteria, which are aggregated to get the overall score. 'Holistic' mark schemes, where assessors decide upon 1 overall score/level, are less common. Being more clearly defined, analytical mark schemes are perhaps more likely to encourage greater consistency, and ensure that assessors are taking each criterion into account (although may not ensure that each element is weighted as was intended). Holistic mark schemes are perhaps less burdensome, but may be less reliable if assessors make decisions in different ways. For example, where a candidate exhibits different levels of performance across different skills, and the assessment criteria are insufficiently precise, it may be difficult for assessors to reliably reconcile those differences when deciding upon a single score (Black & Newton, 2016). For each type of levels-based mark scheme, the number of levels needs to also be considered (either for each criteria, or the single score): too few levels might lead to inadequate discrimination of pupils; too many may make it difficult for assessors to reliably distinguish between them.

When making judgements, examiners and teachers often vary in their adherence to mark schemes or assessment criteria, and often make relative, as opposed to absolute, evaluations of pupils' work (see van Daal, Lesterhuis, Coertjens, Donche, & De Maeyer, 2019). Comparative judgement takes advantage of this fact, building on the idea that it is easier to make relative judgements than absolute judgements, thus potentially improving reliability in judgements (cf Thurstone, 1927). Other advantages include the fact that very little training is needed for assessors compared to other methods, and this approach is able to control for any individual differences in severity/leniency in assessors (see Andrich, 1978; Bramley, 2007). These factors increase the potential for reliability, and indeed good levels of reliability have been reported for assessments of writing using this method (eg Heldsinger & Humphry, 2010, 2013; No More Marking, 2017)[22]. However, the findings reported by Whitehouse (2012) suggested that the shared understanding of quality (in their case, of geography essays) amongst the assessors in their study were based upon existing mark schemes and the training that they had received on those mark schemes as examiners. The question arises then, that if comparative judgement were to be used as the main method of assessment, in the absence of clear marking criteria, whether this shared understanding would be maintained. With less external control, there is a possibility that understanding may differently diverge for each assessor from the construct intended by the assessment developers, raising concerns for both reliability and validity. As with the other methods of marking/grading, the quality of the writing produced may depend on the task set – while decision making may be more reliable under this approach, controls are still needed relating to reliable task setting and the environment in which work is produced.

[22] It should be noted that these studies included materials from across a range of primary school years. This may have improved reliability scores, as it may be easier to discriminate between writing produced by pupils of different ages than between writing of pupils of the same age.

5.3 Conclusions

Assessment is a complex process, requiring a number of different procedures and controls to secure validity. Each of these procedures – for example, setting the assessment, the mode of the assessment, and marking – can be approached differently. Decisions for any assessment design will ultimately depend upon considerations of validity in relation to what the purpose of the assessment is, and what the intended uses of outcomes are (a discussion on the various purposes to which assessments might be put has been presented by Newton, 2007). Other considerations outside the scope of this review would also need to be taken into account, such as feasibility, logistics and cost. Taking each of these factors into account both during and beyond the assessment design stage can help ensure that any assessment of writing offers, and then continues to offer, valid and reliable measurement of this fundamental skill.

Appendix: Tables for the review of international approaches

Table 1. Overview of the identified assessments

Jurisdiction	Assessment	Description	Method of the writing assessment	Targeted pupils	Further information
Australia	National Assessment Program – Literacy and Numeracy (NAPLAN)	Yearly national assessment of reading, writing, spelling and grammar, and numeracy.	Extended response type items. Traditionally paper-based, but a sample of students were tested online in 2018.	All pupils in: Year 3 – Age 8-9 Year 5 – Age 10-11 Year 7 – Age 12-13 Year 9 – Age 14-15	NAP (2016a, 2018) NAPLAN Online (2018) Example test items: NAP (2018)
Canada (Ontario)	Assessment of Reading, Writing and Mathematics: Junior Division (also known as the Junior Division Assessment; JDA)	Yearly provincial assessment of mathematics and language (including reading and writing).	Paper-based, with extended response and multiple-choice type items.	All pupils in: Grade 6 – Age 11-12	EQAO (2007, 2017c) Example tests: EQAO (2017b)
Caribbean Examinations Council (CXC)[a]	Caribbean Primary Exit Assessment (CPEA)	Yearly multi-national assessment of 'common literacies': mathematical, civic and scientific, and language (including writing).	Internally assessed portfolio of collections of writing produced over time.	All pupils in subscribed jurisdictions at the end of primary school – Age 10-11	CXC (n.d., 2016)

England	National curriculum assessments: Key Stage 2 (KS2)	Yearly national assessment of reading, mathematics, grammar punctuation and spelling, and writing.	Internally assessed portfolio of collections of writing produced over time.	All pupils in: Year 6 – Age 10-11	STA (2017b)
Hong Kong	Territory-wide System Assessment (TSA)	Yearly (for Primary 3 and Secondary 3) or biennial (for Primary 6) territory-wide assessment of Chinese and English language (including writing) and mathematics. From 2018 onwards, Primary 3 pupils are assessed on a sampling basis only.	Paper-based, with a single extended response type item.	All pupils in: Primary 3: 8-9 Primary 6: 11-12 Secondary 3: 14-15	HKEAA (n.d.-a, n.d.-b) Example tests: HKEAA (2015a)
New Zealand	Assessment Tools for Teaching and Learning (e-asTTle)	Optional assessment of reading, writing, and mathematics, in English and/or Māori. Can be taken at any point in the year.	Computer-based, with a single extended response type item.	Designed for pupils in: Year 5 – Age 8-9 Year 6 – Age 9-10 Year 7 – Age 10-11 Year 8 – Age 11-12 Year 9 – Age 12-13 Year 10 – Age 13-14	New Zealand Ministry of Education (n.d.-a, n.d.-b)
Pakistan	National Achievement Test (NAT)	Yearly sample-based assessment of maths, science, social	Paper based. Item type is unclear. However, the writing	A random stratified sample of pupils in: Grade 4 – Age 9-10	Pakistan Ministry of Federal

		studies, and reading and writing in Urdu, Sindhi, and English.	items are marked by professional markers standardised to understand the 'spirit' of marking rubrics. This would imply extended-response type items, as short answers or MCQs would have more clearly defined rubrics.	Grade 8 – Age 13-14 50% of all government schools are sampled, and then 20 pupils are sampled from each school.	Education and Professional Training (2016a, 2016b)
Philippines	National Achievement Test (NAT)	Yearly national assessment of science, maths, English (including writing), Filipino, and social studies.	Paper-based, with multiple-choice type items.	All public school pupils in: Grade 3 – Age 8-9 All pupils in: Grade 6 – Age 11-12 2nd Year – Age 13-14	Benito (2010)
Scotland	Scotland National Standardised Assessments (SNSA)	Yearly national assessment of reading, writing, and numeracy.	Computer-based, with mostly multiple-choice type items (word choice, also including matching of items and drag and drop selection) and some single word typed answers.	All pupils in: Primary 1 – Age 4-5 Primary 4 – Age 7-8 Primary 7 – Age 10-11 Secondary 3 – Age 13-14	SNSA (n.d.-a, n.d.-b)
Singapore	Primary School Leaving Examination (PSLE)	Yearly national assessment of writing, language use and comprehension,	Paper-based, with 2 extended response type items.	All pupils in: Primary 6 – Age 11-12	SEAB (2015, 2018a)

		listening comprehension, and oral communication.			
Trinidad & Tobago	Secondary Entrance Assessment (SEA)	Yearly national assessment of writing, mathematics and language (spelling and grammar and reading comprehension).	Paper based, with 1 extended response type item.	All pupils in: Standard 5 – Age 10-11	Republic Of Trinidad & Tobago Ministry Of Education (2017a)
Uganda	Primary Leaving Examinations (PLE)	Yearly national assessment of English language (including writing), mathematics, science, and social studies.	Paper-based, with short response type items (eg single word or single sentence answers).	All pupils in: Primary 7 – Age 11-12	UNEB (2016)
United States of America (California)	English Language Proficiency Assessments for California (ELPAC)	Assessment of proficiency in English for non-native speakers. Includes listening, speaking, reading and writing.	Paper-based, with a mixture of item types: short responses (1 or 2 sentences) and longer extended responses (1 or more paragraphs).	Any non-native speaker of English is tested within 30 days of first enrolment into any class from kindergarten (age 5-6) through to grade 12 (age 17-18). The assessment is then repeated annually throughout school until the pupil is deemed fluent in English.	California Department of Education (n.d.-b, n.d.-c) Example tests: California Department of Education (2018a)

United States of America (California)	California Assessment of Student Performance and Progress (CAASPP)[b]	Yearly state-wide assessment of maths and English (including writing).	Computer-based, with a mixture of multiple-choice, alternative format (eg clicking on sections of text), single paragraph and multiple paragraph extended response type items.	All pupils in: Grade 3 – Age 8-9 Grade 4 – Age 9-10 Grade 5 – Age 10-11 Grade 6 – Age 11-12 Grade 7 – Age 12-13 Grade 8 – Age 13-14 Grade 11 – Age 16-17	CAASPP (n.d., 2016) Example test items: Smarter Balanced Assessment Consortium (n.d.)
United States of America (National)	National Assessment of Educational Progress (NAEP)	National assessment of 10 subjects, including writing. Only some subjects are assessed in each year, with writing being assessed every 4-6 years (the last writing assessment was in 2017, and the next is scheduled for 2021).	Computer-based assessment, with each pupil completing 2 extended-response type items.	A random stratified sample of pupils in: Grade 4 – Age 9-10 Grade 8 – Age 13-14 Grade 12 – Age 17-18 A nationally representative sample of schools is selected with regards to region, ethnic composition, and student achievement. For each school selected, around 10 pupils are sampled per grade per subject.	NCES (2018a, 2018b) NAGB (2017b) Example tests: NCES (2018c)

[a] CXC membership countries include Anguilla, Antigua and Barbuda, Barbados, Belize, British Virgin Islands, Cayman Islands, Dominica, Grenada, Guyana, Jamaica, Montserrat, St. Kitts and Nevis, St. Lucia, St. Vincent and the Grenadines, Trinidad and Tobago and Turks and Caicos Islands

[b] The main component of the CAASPP is also known as ‘Smarter Balanced Summative Assessments’

Table 2. Uses and Stakes/Accountability

Jurisdiction & assessment	**Intended uses**	**Stakes/Accountability**
Australia – NAPLAN Computer-based test	"[The NAPLAN] provides information for parents, teachers, and schools on individual student progress… [to] know how well their students are performing, compared with other children across Australia, and if there are areas where a child needs support and further assistance… [It] also gives… ministers the information they need to provide greater support for schools or students in specific areas or years" (NAP, 2016c).	While not necessarily high-stakes by design, the fact that school results are published online (allowing for comparisons between schools to be made) means that the stakes of the NAPLAN has become raised for many stakeholders (discussed by Thompson, 2013). "[Pupils do] not receive a pass or fail classification from ACARA after a NAPLAN test. However, their results are used for various purposes by jurisdictions and schools including the granting of access to selective educational programs and targeted teaching." (ACARA, personal communication, December 19th, 2018).
Canada (Ontario) – JDA Paper-based test	"The purpose of the [JDA] (Grades 4–6) is to assess the level at which students are meeting curriculum expectations in reading, writing and mathematics at the end of the junior division (up to the end of Grade 6)." (EQAO, 2007, p. 6). In addition to attitudinal and behavioural surveys, the achievement data of the JDA is used to help determine instructional strategies, planning, and resource allocation (EQAO, personal communication, January 11th, 2019).	Results do not count towards individual pupils' grades, however outcomes are used to "strengthen the accountability of the public education system" (EQAO, 2014). Results for each pupil are also sent to each student's parents (EQAO, 2018b). Reports are published which show how schools compare to others in the same school board and province (see EQAO, 2017a), which helps keeps schools accountable (EQAO, personal communication, January 11th, 2019).

Caribbean Examinations Council (CXC) – CPEA Portfolio	"The CPEA will provide the foundation for a seamless transition to secondary education and facilitate portability of qualifications across the Caribbean Region. It will: 1. assist with the quality measures in the primary education system; 2. offer a common measure across schools and territories in the region; 3. respond to the calls for a regional assessment at the primary level." (CXC, 2016, p. 1)	Outcomes are used for teacher/school accountability. Some media reports suggests pressure on schools to maintain their ranking, and on pupils to succeed (eg Searchlight, 2014; Vincentian, 2017).
England – KS2 Portfolio	"The tests help measure the progress pupils have made and identify if they need additional support in a certain area. The tests are also used to assess schools' performance and to produce national performance data." (STA, 2018, p. 3)	Outcomes contribute to primary school accountability measures. Increased pressure/workload for teachers as a result of this accountability has been noted (House of Commons Education Committee, 2017).
Hong Kong – TSA Paper-based test	At the territory-wide level, TSA is used to help inform the government with regards to setting education policies, allocating funding and resources, reviewing the curriculum, and to use data for research. At the school level, TSA is used to help schools understand their pupils' strengths and weaknesses (as a cohort) (HKEAA, personal communication, January 9th, 2019).	"[The] TSA isn't meant to rank schools and students [and] it doesn't affect the allocation of [secondary school] places" (HKEAA, n.d.-a, Video 1). Extra tuition specifically for the TSA is actively discouraged by the assessment authority; it is very much promoted as being a low-stakes test (HKEAA, n.d.-a, n.d.-b). "TSA doesn't assess schools or teachers [and is not a] performance appraisal report" (HKEAA, n.d.-a, Video 6).
New Zealand – e-asTTle	"e-asTTle provides teachers and school leaders with information that can be used to inform learning programmes and to apply teaching	The test is optional, with no pressure for schools to make outcomes publicly available.

Computer-based test	practice that maximises individual student learning. Schools using e-asTTle have found it to be a great tool for planning purposes, for helping students to understand their progress, and for involving parents in discussions about how well their children are doing." (New Zealand Ministry of Education, n.d.-a). "The Ministry [of Education's] use of information in the e-asTTle dataset is predominantly for research purposes… [for example] to develop insights on how students of different characteristics are progressing… The outputs of this type of research will feed into internal policy discussions, as well as being published."	"The Ministry [of Education] does not use e-asTTle data to identify the performance of individual students, schools or teachers." (New Zealand Ministry of Education, personal communication, January 8th, 2019)
Pakistan – NAT Paper-based test	The objectives of the [NAT] are to inform policymakers on the correlations between geography/gender and performance, to monitor standards, to inform funding allocation, and to provide information for teachers on how to improve student performance (Pakistan Ministry of Federal Education and Professional Training, 2016b).	"[The NAT] is 'low-stakes'… for individuals (but 'high-stakes for the nation)… It is not an assessment of individuals – whether students, teachers, or schools; or a 'high-stakes' examination… for promotion or selection" (Pakistan Ministry of Federal Education and Professional Training, 2016b).
Philippines – NAT Paper-based test	"The test aims to: 1) provide empirical information on the achievement level of pupils… to serve as guide for policy makers, administrators, curriculum planners, supervisors, principals and teachers in their respective courses of action; 2) identify and analyze variations on achievement levels across the	Schools are held accountable on their outcomes (Benito, 2010), so the test would seem to be high stakes.

	years by region, division, school and other variables; 3) determine the rate of improvement in basic education" (Benito, 2010, p. 7).	
Scotland – SNSA Computer-based test	"At a classroom level, the information provided from children and young people's assessments will help teachers to understand how children are progressing in… reading, writing and numeracy… School level data will be available to teachers and local authorities to help them tailor their own improvement planning. Scottish Government will have access to national level data only. This is to help identify trends, drive national policy and improvement priorities… and will, in turn, inform the type and level of national support required." (SNSA, n.d.-a) The assessments are designed to be diagnostic and formative in that they can support planning of next steps in learning when considered alongside wider assessment evidence (SNSA, personal communication, January 29th, 2019).	"The assessments are not 'high stakes tests'. The results do not determine any key future outcomes for students, such as which school they go to, or whether they can progress to the next level. There is no pass or fail. Children are not expected to revise or prepare for assessments. They will simply continue to undertake routine classroom learning activities." (SNSA, n.d.-a)
Singapore – PSLE Paper-based test	"The PSLE assesses how much students have learnt over six years of primary education and whether they have acquired a sufficient academic foundation to access and benefit from secondary education. It also helps students, parents and teachers to determine where each child's strengths lie and tailor secondary school education to best suit a child's learning needs, so that students can receive the necessary support." (Singapore Ministry of Education, 2016a).	Various efforts have been taken to lower the stakes of the tests. These include reduced reporting of outcomes and introducing wider grade bands to discourage pupil rankings and stress based on small mark differences. (Singapore National Library Board, 2016). Nevertheless, outcomes still affect progression, and so do remain high stakes (Singapore Ministry of Education, 2016b, and SEAB, personal communication, December 31st, 2018).

Trinidad & Tobago – SEA Paper-based test	"The Secondary Entrance Assessment (SEA) Examination is used to facilitate the placement of students in Secondary Schools throughout Trinidad and Tobago." (Government of the Republic of Trinidad and Tobago, 2018)	Scores have a direct impact on secondary school placements for pupils (Republic Of Trinidad & Tobago Ministry Of Education, 2017b).
Uganda – PLE Paper-based test	"A student's score on the government test at end of primary 7 serves as kind of "admission" test for private schools and some government schools (which are not to be confused with USE schools.)" (LaMendola, 2014)	Outcomes affect pupil progression to secondary school. Competition also seems to exist between primary schools in securing the best outcomes (The Observer, 2018).
United States of America (California) – ELPAC Paper-based test	"Identifying students who need help learning in English is important so these students can get the extra help they need to do well in school and access the full curriculum" (California Department of Education, n.d.-c).	Outcomes do not seem to be high stakes for either individuals or for schools.
United States of America (California) – CAASPP Computer-based test	"[The CAASPP can] help facilitate conversations between parents/guardians and teachers about student performance; Serve as a tool to help parents/guardians and teachers work together to improve student learning; Help schools and school districts identify strengths and areas that need improvement in their educational programs; Provide the public and policymakers with information about student achievement" (CAASPP, 2016, p. 5).	The CAASPP does not seem to be high stakes for pupils, particularly those in primary education, however schools are required to report outcome data in their School Accountability Report Card (SARC), which is made publicly available online (California Department of Education, n.d.-a).
United States of America (National) – NAEP	"[The NAEP] provides the only national report on student achievement in a variety of subjects… [It] includes information on the performance of	"NAEP is designed to produce group scores, and is prohibited by Congress from reporting individual student results… By law, NAEP is forbidden to

Computer-based test	various subgroups of students at the national, state, and urban district levels." (NAGB, 2017b, p. 43). "Because states and local education agencies determine their own curriculum and … student assessments within their respective jurisdictions, NAEP is used by local and national policymakers to understand the achievement of various student groups. [It gives] a clear sense of the degree to which that nation as a whole is meeting its goals in student writing achievement." (U.S. Department of Education, personal communication, December 31st, 2018) Educators, policymakers, and elected officials all use NAEP results to develop ways to improve education (NCES, 2017).	report individual school results, to influence curriculum, and to be used for high-stakes purposes" (NAGB, 2017a).

Table 3. The writing component: objectives, tasks, and marking/grading

Jurisdiction & Test	Assessment objectives	Tasks	Marking and grading
Australia – NAPLAN High-stakes Computer-based test	The NAPLAN focusses on narrative writing and persuasive writing, although only 1 type is assessed each year. Other skills are assessed (eg structure and grammar, punctuation, and spelling), but the assessment mainly focuses on 'purpose and audience' (ie compositional skills) (NAP, 2018, and ACARA, personal communication, December 19th, 2018).	Students are asked to respond to a given prompt (an idea or a topic). For example, they might be asked to write an opinion piece (persuasive writing) or a story (narrative writing). Pupils do not know beforehand which type of writing they will be asked to produce (NAP, 2018). The prompt/genre for each year is chosen according to item performance on pre-test trials with over 1000 pupils (eg by considering psychometric analyses, marker feedback, word counts, and accessibility). Pupils/teachers are not made aware which genre will be tested prior to the main test (ACARA, personal communication, December 19th, 2018). Students are given 5 minutes to plan, 30 minutes to write, and 5 minutes to edit their response (NAP, 2010, 2013).	Submissions are externally marked, using levels-based mark schemes, supported by level descriptors, exemplars, and a glossary of terms. Separate (but only marginally different) mark schemes exist for the narrative and persuasive tasks, but the same marks schemes are used for all year groups. Separate marks are awarded for 10 different marking criteria, which relate to a mixture of technical skills (eg spelling and punctuation) and compositional skills (eg engaging and persuading the audience), presented in manner largely specific to the type of writing being assessed (NAP, 2010, 2013). Marks are transformed into 1 of 10 'bands'. All year groups are placed onto the same scale, and bands are divided into working 'below / at / above the national minimum standard', with the position of these divisions differing by year group (eg Band 2 is the national minimum standard for Year 3, but Band 6 is the national minimum standard for Year 9 – see NAP, 2016b).
Canada (Ontario) – JDA	The writing element of the JDA focusses on 3 writing skills:	The language paper is divided into 4 sections: Section A contains a short writing prompt and 5 writing multiple-	The extended writing elements are scored by assigning them to one of 6 'codes' (similar to levels) for 'topic development' (how developed

High-stakes Paper-based test	"developing a main idea with sufficient supporting details", "organizing information and ideas in a coherent manner", and "using conventions (spelling, grammar, punctuation)" (EQAO, 2007, p. 11).	choice items; Section B contains a short writing prompt; Section C contains a short writing prompt and 4 multiple-choice questions; Section D contains a long writing prompt. In addition to these writing tasks, each section also contains items relating to reading. Pupils are allotted 1 hour for each section, but may take more time, so long as each section is done in one continuous sitting (EQAO, personal communication, January 15th, 2019). The short writing prompts require a 1-page response; the long writing prompt requires a 2-page response (EQAO, 2007). A dictionary and thesaurus are allowed for the writing tasks (EQAO, 2018a).	and focussed the response is) and 5 'codes' for 'conventions' (SPAG) (EQAO, 2007). Multiple-choice items are machine marked (EQAO, 2007). Pupils are assigned 1 of 5 overall outcomes. Levels 1-4 reflect achievement that ranges from "much below the provincial standard" to "surpasses the provincial standard". Any who fail to meet Level 1 are assigned an "NE 1" (not enough evidence) (EQAO, 2017a). The longer open-response task is allocated 7 score points (24% of the writing assessment), the shorter open-response tasks are allocated 7 points each (48% of the assessment combined), and the multiple-choice items are allocated 1 score points each (28% of the assessment combined) (EQAO, 2007, and personal communication, January 11th, 2019).
Caribbean Examinations Council (CXC) – CPEA High-stakes Portfolio	Learning outcomes focus on being able to write for a range of different purposes, and showing good levels of organisation and	Pupils produce a writing portfolio (weighted less than 7% of the overall CPEA). Collections of writing are used to show progress over time (CXC, 2016).	Teacher assessment is used, with some self-assessment (ie by the pupil). Information on mark schemes could not be found. However, because mark schemes should be submitted as part of the portfolios, this would suggest that they are developed separately by the teacher (CXC, 2016).

	technical control in writing (CXC, n.d.).		Pupils are ultimately assigned 1 of 4 outcomes, from Level 1 (“Needs improvement”) to Level 4 (“Exemplary”) (CXC, n.d.).
England – KS2 High-stakes Portfolio	The National Curriculum focuses on 2 main elements: ‘transcription (spelling and handwriting)’, and ‘composition (articulating ideas)’ (DfE, 2013).	Pupils produce a writing portfolio of examples of writing throughout Key Stage 2, covering a range of different genres and styles. Writing used in the assessment is expected to be independent of heavy teacher guidance (STA, 2017b).	Pupils are marked according to a number of ‘pupil-can’ statements, and are given an outcome of ‘working towards’, ‘working at’, or ‘working at greater depth [than]’ the expected standard. To achieve each of these standards, pupils must demonstrate that they meet **all** the statements within that standard (a few exceptions apply). A statement is considered ‘met’ when sufficient evidence has been found within their portfolio of writing. Portfolios are internally assessed by teachers following these criteria, a sample of which are externally moderated (STA, 2016).
Hong Kong – TSA Low-stakes Paper-based test	The TSA assesses ‘basic competencies’, which for writing focus upon both technical (eg punctuation, sentence structure) and compositional skills (eg presenting ideas) (HKEAA, n.d.-b, sec. 4 - Key Stage 2). These basic competencies (and therefore the TSA itself) only cover part of the overall	Pupils are assigned 1 of several sub-papers, meaning that they do not all answer the same questions. Pupils are asked to produce an extended piece of writing of about 80 words (eg a story or a letter), based on a given prompt, with about 25 minutes being allocated for this task (HKEAA, 2015a).	Mark schemes are levels-based, supported by level descriptors. Pupils are marked out of 4 for each domain: content (level of detail and clarity) and language (eg vocabulary, verb forms, grammar) (HKEAA, 2015a). Marks are not aggregated to form an overall mark/grade for each pupil – results are summarised at group level for schools and the territory overall (HKEAA, 2015b).

	curriculum for writing (HKEAA, personal communication, January 9th, 2019).		
New Zealand – e-asTTle Low-stakes Computer-based test	The writing element pf the e-asTTle assesses pupil's ability to write for a range of purposes (describe, explain, recount, narrate, persuade) (New Zealand Ministry of Education, 2012, sec. 1.1).	The e-asTTle is an online test which can be taken at any time. 20 prompts are available, covering the 5 writing purposes (describe, explain, recount, narrate, persuade), from which teachers choose 1. Pupils produce a piece of extended writing in response to this prompt, with a time limit of 40 minutes, and also answer a series of questions assessing their attitudes towards writing (New Zealand Ministry of Education, 2012).	The same levels-based marking rubric is used for all types of writing, supported by level descriptors and annotated exemplars. Pupils are marked separately on 7 domains: ideas, structure and language, organisation, vocabulary, sentence structure, punctuation, and spelling. Raw scores are transformed onto a uniform scale (to take into account differences in difficulty between these different domains), and are reported with confidence limits, to recognise measurement error (New Zealand Ministry of Education, 2012).
Pakistan – NAT Low-stakes Paper-based test	The NAT focusses on narrative, informative, and persuasive writing (Pakistan Ministry of Federal Education and Professional Training, 2016a).	The nature of the marking would suggest that pupils produce extended-responses. Further information could not be found.	Tests are externally marked according to rubrics. It is unclear what these rubrics look like, however, the fact that markers are trained to understand the 'spirit' of the rubrics would suggest some form of levels based marking (ie as opposed to specific points-based criteria) (Pakistan Ministry of Federal Education and Professional Training, 2016a).
Philippines – NAT	The English component targets 3 writing competencies: "identify cohesive devices"; "identify	The test comprises of multiple-choice items, most of which are of 'moderate difficulty' (Benito, 2010).	Method of marking is unclear, but they are likely externally marked. Test scores are reported as raw scores and as percentages (Benito, 2010).

High-stakes Paper-based test	correct bibliographic entry"; "fill out forms" (Benito, 2010, p. 17)		
Scotland – SNSA Low-stakes Computer-based test	Writing questions target spelling, grammar, and punctuation only (EIS, 2018).	There is no specific assessment window, and no time limit (but should take less than 45 minutes SNSA, n.d.). Writing items are a combination of multiple-choice (word choice, also including matching of items and drag and drop selection) and single word typed answers (eg spelling). Assessments are adaptive, meaning item demand is adapted according to performance on early items (SNSA, n.d.-b).	Assessments are marked automatically online (SNSA, n.d.-a). Reports are produced for each pupil, which show the number of correct responses, as well as their position on an 'overall capacity demonstrated' scale (from low to medium to high) (SNSA, n.d.-b).
Singapore – PSLE High-stakes Paper-based test	The writing component targets 5 assessment objectives: 1) writing to suit purpose, audience and context; 2) Using appropriate register and tone; 3) Organising and expressing ideas; 4) Spelling and grammar; 5) Vocabulary (SEAB, 2015).	The test lasts 70 minutes, and is weighted 27.5% of the overall PSLE English Language Marks. Pupils are asked to produce 1 piece of 'situational writing' (15 marks), which constitutes a short 'functional piece', such as a letter, email, or report, and 1 piece of 'continuous writing' (40 marks), which constitutes a longer (150 words minimum) piece of continuous prose based upon a given prompt (3 pictures offering "different angles of interpretation") (SEAB, 2015, p. 5).	Tests are externally marked according to a levels-based mark scheme, supported by level descriptors. Pupils are marked according to 2 domains: 'content' and 'language and organisation' (SEAB, personal communication, December 31st, 2019). Pupils are awarded a grade of E to A*, as well as a scaled 'T-score' for each subject, which indicates a pupil's performance relative to that pupil's peers. (SEAB, 2018b) From 2021, the scoring system is changing towards wider bands, which will not be dependent upon peer performance, to replace

			specific T-scores (Singapore Ministry of Education, 2016a).
Trinidad & Tobago – SEA High-stakes Paper-based test	The writing paper focuses on either narrative writing or expository writing (the type of writing may change each year) (Republic Of Trinidad & Tobago Ministry Of Education, 2017a).	Test papers contain either 3 narrative (story) items, or 3 expository (explanatory) items (the type of task assessed may change each year). Pupils choose to write about 1 topic, with a 50 minute time limit (Republic Of Trinidad & Tobago Ministry Of Education, 2017a).	Pupils are externally double-marked (ie by 2 examiners) on content, language use, grammar and mechanics, and organisation (Republic Of Trinidad & Tobago Ministry Of Education, 2017a). The 'holistic marking' approach would suggest a levels-based mark scheme (Republic Of Trinidad & Tobago Ministry Of Education, 2004). Raw scores are reported alongside a scaled overall score and percentile rank of the pupil nationally (Republic Of Trinidad & Tobago Ministry Of Education, 2017b).
Uganda – PLE High-stakes Paper-based test	Information not found.	An official source could not be found, but some revision materials were. The writing element of the test seems to mostly comprise of a series of items requiring the writing of single word or single sentence answers. These focus on word choice/usage, vocabulary, and sentence structure (ReviseNow, 2018) .	Tests are externally marked (Daily Monitor, 2016). The exact method of marking is unclear. For the English component overall, pupils are awarded a fail to distinction (UNEB, 2016).
United States of America (California) – ELPAC	The ELPAC covers writing of "literary and informational texts to present, describe, and explain ideas and information in a range	Pupils complete a number of tasks. There are 4 short response type items (2 relating to 'describe a picture', and 2 relating to 'write about academic information') and 2 extended response type items ('write	Responses are externally marked using levels-based mark schemes, supported by level descriptors. Different mark schemes exist for each type of task, so that there are focussed criteria on describing a picture, writing about

Low-stakes Paper-based test	of social and academic contexts" (California Department of Education, 2018a, p. 24).	about an experience' and 'justify an opinion') (California Department of Education, 2018a). There is no time limit for the test, but 40-50 minutes has been suggested (for Grades 3-5) (California Department of Education, 2018b).	academic information or experiences, or justifying opinions (ETS, 2018). Pupils are assigned 1 of 4 levels based upon their outcomes, from level 1 ("minimally developed") to level 4 ("well developed") (California Department of Education, n.d.-d).
United States of America (California) – CAASPP High-stakes Computer-based test	The writing standards of the Common Core State Standards (California Department of Education, 2013) for kindergarten to Grade 5 (up to age 11) focus on writing for different purposes, writing coherently and editing text, producing research, and writing both over short timeframes and extended timeframes. There is a focus on opinion-based, informative, and narrative writing.	Pupils complete a test and a performance task, which are both completed online. The test contains a mixture of multiple-choice, alternate format items (clicking on sections of text) and extended response type items (single paragraph) (Smarter Balanced Assessment Consortium, 2017). The test is adaptive, meaning that the demands of questions presented to pupils change according to performance on earlier questions (CAASPP, 2016). The performance task contains a mixture of multiple-choice and 3 extended response type items (2 single paragraphs, and 1 multiple paragraphs) (Smarter Balanced Assessment Consortium, 2014).	Responses are externally marked. Multiple-choice and alternate format items are likely auto-marked (as part of the adaptive nature of the test). Extended responses are marked according to levels-based mark schemes, supported by level descriptors. Marking criteria focus mainly on writing for a purpose (eg developing narrative, presenting evidence), with a limited focus on technical writing skills (Smarter Balanced Assessment Consortium, 2014, 2017). Pupils' scores for the overall English component are converted into 1 of 4 achievement levels: 'standard not met', 'standard nearly met', 'standard met', 'standard exceeded' (CAASPP, 2016).
United States of America	The NAEP evaluates "writers' ability… To Persuade; To Explain;	Pupils sampled for the writing test complete 1 of several different test booklets, with each pupil completing	Responses are externally evaluated on 3 features: 'development of ideas', 'organisation of ideas', and 'language facility and

(National) – NAEP Low-stakes Computer-based test	and To Convey Experience, Real or Imagined. Because understanding the nature of one's audience is fundamental to successful communication, writing tasks will specify or clearly imply an audience, and writers will be asked to use approaches that effectively address that audience." (NAGB, 2017b, p. vi)	a subset of items. Booklets are distributed so that each area (persuading, explaining, conveying experience) is covered by a representative sample of students (NCES, 2018b). Each pupil completes 2, 30-minutes extended-response type tasks (which could be on any of the 3 areas, though no student addresses the same area twice), in response to a given prompt (some tasks include multimedia stimuli). In each task, the intended audience of the writing is clearly stated/implied (NAGB, 2017b). Alongside the test, pupils also complete a questionnaire designed to gather information on 'contextual variables', such as learning habits and attitudes (NAGB, 2017b).	conventions' – these features are evaluated in relation to the stated purpose and audience of each task. A levels-based holistic marking scheme, supported by level descriptors, is used to give each pupil a single score of 1-6 for each task (ie rather than assessing each of the 3 features separately, before producing an aggregated score). A separate mark scheme is used for each purpose of writing (persuading, explaining, or conveying experience). Scores are also reported on 3 levels of achievement: 'basic', 'proficient', and 'advanced'. (NAGB, 2017b)

References

ACARA. (2015). *An Evaluation of Automated Scoring of Naplan Persuavie Writing*. Sydney, Australia: Australian Curriculum, Assessment and Reporting Authority. Retrieved from http://nap.edu.au/_resources/20151130_ACARA_research_paper_on_online_automated_scoring.pdf

Ahmed, A., & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy & Practice*, *18*, 259–278. http://doi.org/10.1080/0969594X.2010.546775

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, *2*, 451–462. http://doi.org/10.1177/014662167800200319

Benito, N. V. (2010). *National Achievement Test: An overview*. Retrieved from http://www.cfo-pso.org.ph/pdf/9thconferencepresentation/day2/National_Achievement_Test_Dr Benito.pdf

Bew, P. (2011). *Independent Review of Key Stage 2 testing, assessment and accountability*. London, UK: Department for Education. Retrieved from https://www.gov.uk/government/publications/independent-review-of-key-stage-2-testing-assessment-and-accountability-final-report

Black, B., & Bramley, T. (2008). Investigating a judgemental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education*, *23*, 357–373. http://doi.org/10.1080/02671520701755440

Black, B., & Newton, P. E. (2016, November). *Tolerating difference of opinion*. Paper presented at the 17th Annual AEA-Europe Conference. Cyprus.

Bramley, T. (2007). Paired comparison methods. In P. E. Newton, J.-A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards2* (pp. 246–294). London, UK: Qualifications and Curriculum Authority.

Bramley, T. (2015). *Investigating the reliability of Adaptive Comparative Judgment*. Cambridge, UK: Cambridge Assessment.

Bramley, T., & Crisp, V. (2019). Spoilt for choice? Issues around the use and comparability of optional exam questions. *Assessment in Education: Principles, Policy & Practice*, *26*, 75–90. http://doi.org/10.1080/0969594X.2017.1287662

Bramley, T., & Vitello, S. (2019). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, *26*, 43–58. http://doi.org/10.1080/0969594X.2017.1418734

CAASPP. (n.d.). Understanding Scores: Grade Six English Language Arts/Literacy. Retrieved June 19, 2018, from http://testscoreguide.org/ca/scores/grade6ela.html?range=met

CAASPP. (2016). *Parent guide to the smarter balanced summative assessments*. Sacramento, CA: California Department of Education. Retrieved from http://testscoreguide.org/ca/resources/

California Department of Education. (n.d.-a). A parent's guide to the SARC. Retrieved June 19, 2018, from https://www.cde.ca.gov/ta/ac/sa/parentguide.asp

California Department of Education. (n.d.-b). English Language Proficiency Assessments for California (ELPAC). Retrieved June 21, 2018, from https://www.cde.ca.gov/ta/tg/ep/

California Department of Education. (n.d.-c). Parent Guide to Understanding the ELPAC. Retrieved June 21, 2018, from https://www.cde.ca.gov/ta/tg/ep/elpacparentguide.asp

California Department of Education. (n.d.-d). Summative ELPAC General PLDs. Retrieved June 21, 2018, from https://www.cde.ca.gov/ta/tg/ep/elpacgpld.asp

California Department of Education. (2013). *California Common Core State Standards*. California Department of Education. Retrieved from https://www.cde.ca.gov/re/cc/

California Department of Education. (2018a). *ELPAC - Practice Test: Grades 3-5*. California Department of Education. Retrieved from https://www.elpac.org/resources/practicetests/

California Department of Education. (2018b). English Language Proficiency Assessments for California (ELPAC) Spring 2018 Summative Assessment: Revised Estimated Testing Times. Retrieved July 20, 2018, from https://www.elpac.org/test-administration/sa-estimated-test-time/

Cooksey, R. W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as judgment-in-context: Analysing how teachers evaluate students' writing. *Educational Research and Evaluation*, *13*(5), 401–434. http://doi.org/10.1080/13803610701728311

Cresswell, M. J., & Houston, J. G. (1991). Assessment of the National Curriculum — some fundamental considerations. *Educational Review*, *43*, 63–78. http://doi.org/10.1080/0013191910430106

Cuff, B. M. P. (2017). *An exploratory investigation into how moderators of non-examined assessments make their judgements*. (Ofqual Report 17/6252). Coventry, UK: Office of Qualifications and Examinations Regulation. Retrieved from https://www.gov.uk/government/publications/exam-and-assessment-marking-research

Cuff, B. M. P. (2018). *International approaches to the moderation of non-examination assessments in secondary education*. (Ofqual Report 18/6364). Coventry, UK: Office of Qualifications and Examinations Regulation. Retrieved from https://www.gov.uk/government/publications/international-literature-review-of-secondary-assessments

Cuff, B. M. P., Howard, E., Mead, R., & Newton, P. E. (2018). *Key stage 2 writing moderation: Observations on the consistency of moderator judgements*. (Ofqual Report 18/6358). Coventry, UK: Office of Qualifications and Examinations Regulation. Retrieved from https://www.gov.uk/government/publications/observations-on-the-consistency-of-moderator-judgements

CXC. (n.d.). *Caribbean Primary Exit Assessment: Book Reports, Writing Portfolios*. Caribbean Examinations Council. Retrieved from

http://www.cxc.org/examinations/cpea/manuals/

CXC. (2016). *Caribbean Primary Exit Assessment: Handbook for Administrators*. Kingston, Jamaica: Caribbean Examinations Council. Retrieved from http://www.cxc.org/examinations/cpea/

Daily Monitor. (2016). 16,000 examiners to mark PLE, UCE as UNEB sets new UACE grading. Retrieved July 20, 2018, from http://www.monitor.co.ug/News/Education/16-000-examiners-mark-PLE-UCE-UNEB-UACE-grading--guidelines-/688336-3477766-usbidd/index.html

Daugherty, R. (1995). *National Curriculum Assessment: A review of policy 1987-1994*. London, UK: The Falmer Press.

Dearing, R. (1994). *The National Curriculum and its Assessment: Final Report*. London, UK: School Curriculum and Assessment Authority. Retrieved from http://www.educationengland.org.uk/documents/dearing1994/dearing1994.html

DfE. (2013). *The national curriculum in England: Key stages 1 and 2 framework document*. London, UK: Department for Education. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/425601/PRIMARY_national_curriculum.pdf

DfE. (2016). Primary school accountability. Retrieved July 10, 2017, from https://www.gov.uk/government/publications/primary-school-accountability

EIS. (2018). Scottish National Standardised Assessments - Update. Retrieved August 7, 2018, from https://www.eis.org.uk/Education-Updates/SNSA

EQAO. (2007). *Framework: Assessment of Reading, Writing and Mathematics, Junior Division (Grades 4-6)*. Toronto, Canada: Education Quality and Accountability Office. Retrieved from http://www.eqao.com/en/assessments/junior-division/educators/Pages/educators.aspx

EQAO. (2014). *Guide to EQAO assessments*. Toronto, Canada: Education Quality and Accountability Office. Retrieved from http://www.eqao.com/en/assessments/communication-docs/guide-elementary-assessments-english.pdf

EQAO. (2017a). *EQAO's assessment of reading, writing and mathematics, Junior Division (Grades 4-6): Individual Student Report, 2017*. Toronto, Canada: Education Quality and Accountability Office. Retrieved from http://www.eqao.com/en/assessments/junior-division/educators/Pages/educators.aspx

EQAO. (2017b). Examples of the Assessments: Grade 6, Junior Division, 2018. Retrieved March 5, 2019, from http://www.eqao.com/en/assessments/junior-division/Pages/example-assessment-materials-current-year.aspx

EQAO. (2017c). Grade 6 Junior Division Assessment. Retrieved June 4, 2018, from http://www.eqao.com/en/assessments/junior-division/Pages/junior-division.aspx

EQAO. (2018a). *Administration and accommodation guide*. Toronto, Canada: Education Quality and Accountability Office. Retrieved from http://www.eqao.com/en/assessments/junior-division/educators/Pages/educators.aspx

EQAO. (2018b). *Guide to EQAO Assessments in Elementary School*. Education Quality and Accountability Office: Toronto, ON, Canada. Retrieved from http://www.eqao.com/en/assessments/junior-division/parents/Pages/parents.aspx

ETS. (2018). *Writing Rubrics for the English Language Proficiency Assessments for California*. Princeton, NJ: Educational Testing Service. Retrieved from https://www.elpac.org/resources/practicetests/

Flower, L. (1979). Writer-Based Prose: A Cognitive Basis for Problems in Writing. *College English*, *41*, 19. http://doi.org/10.2307/376357

Gove, M. (2013). Education reform: new national curriculum for schools. Retrieved July 5, 2018, from https://www.gov.uk/government/speeches/education-reform-new-national-curriculum-for-schools

Gove, M. (2014). The purpose of our school reforms. Retrieved July 5, 2018, from https://www.gov.uk/government/speeches/the-purpose-of-our-school-reforms

Government of the Republic of Trinidad and Tobago. (2018). Secondary Entrance Assessment (SEA). Retrieved June 4, 2018, from https://www.ttconnect.gov.tt/gortt/portal/ttconnect/Cit_studentDetail/?WCM_GLOBAL_CONTEXT=/gortt/wcm/connect/GorTT+Web+Content/TTConnect/Citizen/Role/AStudent/Examinations/Secondary+Entrance+Assessment+%28SEA%29

Hall, K., & Harding, A. (2002). Level descriptions and teacher assessment in England: towards a community of assessment practice. *Educational Research*, *44*, 1–16. http://doi.org/10.1080/00131880110081071

Harlen, W. (2004). *A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes*. London, UK: EPPI-Centre. Retrieved from http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=116

Harlen, W. (2007). *The Quality of Learning: Assessment Alternatives for Primary Education*. Cambridge, UK: University of Cambridge Faculty of Education.

He, Q., Anwyll, S., Glanville, M., & Deavall, A. (2013). An investigation of the reliability of marking of the Key Stage 2 National Curriculum English writing tests in England. *Educational Research*, *55*, 393–410. http://doi.org/10.1080/00131881.2013.844942

Heldsinger, S., & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher*, *37*, 1–19. http://doi.org/10.1007/BF03216919

Heldsinger, S., & Humphry, S. (2013). Using calibrated exemplars in the teacher-assessment of writing: an empirical study. *Educational Research*, *55*, 219–235. http://doi.org/10.1080/00131881.2013.825159

HKEAA. (n.d.-a). TSA. Retrieved June 5, 2018, from www.bca.hkeaa.edu.hk/web/TSA/en/Introduction.html

HKEAA. (n.d.-b). TSA: Frequently asked questions. Retrieved June 5, 2018, from www.bca.hkeaa.edu.hk/web/TSA/en/Faq.html#Q01

HKEAA. (2015a). TSA: Question papers and marking schemes. Retrieved June 5, 2018, from www.bca.hkeaa.edu.hk/web/TSA/en/PriPaperSchema.html

HKEAA. (2015b). TSA Report. Retrieved June 5, 2018, from www.bca.hkeaa.edu.hk/web/TSA/en/2015tsaReport/priSubject_report_eng.html

House of Commons Education Committee. (2017). *Primary assessment: Eleventh Report of Session 2016-17*. London, UK: House of Commons. Retrieved from https://www.parliament.uk/business/committees/committees-a-z/commons-select/education-committee/inquiries/parliament-2015/primary-assessment-16-17/

Koretz, D. M. (1998). Large-scale Portfolio Assessments in the US: evidence pertaining to the quality of measurement. *Assessment in Education: Principles, Policy & Practice*, *5*, 309–334. http://doi.org/10.1080/0969595980050302

LaMendola, S. (2014). *Learning from Uganda: Reflections and Observations*. Wilkes-Barre, PA: King's College. Retrieved from https://www.kings.edu/non_cms/pdf/Ugandan-Education-System.pdf

Lazendic, G., Justus, J.-A., & Rabinowitz, S. (2018). *Naplan Online Automated Scoring Research Program: Research Report*. Sydney, Australia: Australian Curriculum, Assessment and Reporting Authority. Retrieved from http://nap.edu.au/online-assessment/research-and-development/automated-essay-scoring

McGrane, J., Chan, J., Boggs, J., Stiff, J., & Hopfenbeck, T. N. (2018). *The assessment and moderation of Primary-level writing in contexts where English is the main language of instruction: A systematic review commissioned by Oxford University Press*. Oxford, UK: Oxford University Press.

Moffett, J. (1983). *Teaching the universe of discourse*. Boston, MA: Houghton Mifflin.

NAGB. (2017a). What Is NAEP? Retrieved September 27, 2018, from https://www.nagb.gov/about-naep/what-is-naep.html

NAGB. (2017b). *Writing framework for the 2017 National Assessment of Educational Progress*. Washington, DC, USA: National Assessment Governing Board. Retrieved from https://www.nagb.gov/naep-frameworks/writing/2017-writing-framework.html

NAP. (2010). *Writing: Narrative Marking Guide*. Sydney, Australia: Australian Curriculum, Assessment and Reporting Authority. Retrieved from http://www.nap.edu.au/naplan/writing

NAP. (2013). *Persuasive Writing Marking Guide*. Sydney, Australia: Australian Curriculum, Assessment and Reporting Authority. Retrieved from http://www.nap.edu.au/naplan/writing

NAP. (2016a). FAQs. Retrieved June 4, 2018, from http://www.nap.edu.au/information/faqs/naplan--writing-test

NAP. (2016b). How to interpret. Retrieved June 4, 2018, from http://www.nap.edu.au/results-and-reports/how-to-interpret

NAP. (2016c). National Literacy Numeracy Assessment. Retrieved June 5, 2018, from http://www.nap.edu.au/results-and-reports/how-to-interpret/national-literacy-numeracy-assessment

NAP. (2018). Writing. Retrieved June 4, 2018, from http://www.nap.edu.au/naplan/writing

NAPLAN Online. (2018). *Frequently asked questions: NAPLAN Writing Test*. Retrieved from http://www.nap.edu.au/docs/default-source/default-document-library/naplan-writing-test-faq.pdf?sfvrsn=2

National Education Union. (2016). Primary schools in chaos and despair with new KS1 and KS2 assessments. Retrieved July 25, 2018, from https://www.teachers.org.uk/news-events/press-releases-england/primary-schools-chaos-and-despair-new-ks1and-ks2-assessments

NCES. (2017). Participating in NAEP?: Why Your Participation Matters. Retrieved September 26, 2018, from https://nces.ed.gov/nationsreportcard/participating/

NCES. (2018a). About NAEP. Retrieved September 26, 2018, from https://nces.ed.gov/nationsreportcard/about/

NCES. (2018b). Participant Selection. Retrieved September 26, 2018, from https://nces.ed.gov/nationsreportcard/assessment_process/selection.aspx

NCES. (2018c). Sample Questions Booklets. Retrieved March 5, 2019, from https://nces.ed.gov/nationsreportcard/about/booklets.aspx

New Zealand Ministry of Education. (n.d.-a). e-asTTle basics. Retrieved June 15, 2018, from https://e-asttle.tki.org.nz/About-e-asTTle/Basics

New Zealand Ministry of Education. (n.d.-b). e-asTTle features. Retrieved June 15, 2018, from https://e-asttle.tki.org.nz/About-e-asTTle/Features

New Zealand Ministry of Education. (2012). *e-asTTle writing (revised) manual*. Retrieved from https://e-asttle.tki.org.nz/Teacher-resources#manual

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice*, *14*, 149–170. http://doi.org/10.1080/09695940701478321

No More Marking. (2017). Sharing Standards Results: Year 3. Retrieved July 31, 2018, from https://blog.nomoremarking.com/sharing-standards-results-year-3-286e6a4288df

Odell, L. (1981). Defining and assessing competence in writing. In C. R. Cooper (Ed.), *The Nature and Measurement of Competency in English*. Urbana, US: National Council of Teachers of English.

Pakistan Ministry of Federal Education and Professional Training. (2016a). *National Assessment Report 2016*. Islamabad, Pakistan: Ministry of Federal Education & Professional Training. Retrieved from http://www.neas.gov.pk/Document Center.html

Pakistan Ministry of Federal Education and Professional Training. (2016b). *NEAS brochure*. Retrieved from http://www.neas.gov.pk/Document Center.html

Perelman, L. (2017). *Automated Essay Scoring and NAPLAN: A Summary Report*. Retrieved from https://www.nswtf.org.au/files/automated_essay_scoring_and_naplan.pdf

Pollitt, A. (2012a). Comparative judgement for assessment. *International Journal of Technology and Design Education*, *22*, 157–170. http://doi.org/10.1007/s10798-011-9189-x

Pollitt, A. (2012b). The method of Adaptive Comparative Judgement. *Assessment in*

Education: Principles, Policy & Practice, *19*, 281–300. http://doi.org/10.1080/0969594X.2012.665354

QCDA. (2010). *Assessment and reporting arrangements*. Coventry, UK: Qualifications and Curriculum Development Agency. Retrieved from http://dera.ioe.ac.uk/14734/

Republic Of Trinidad & Tobago Ministry Of Education. (2004). *Secondary Entrance Assessment Guidelines*. Retrieved from https://www.ttconnect.gov.tt/gortt/portal/ttconnect/!ut/p/a1/jdDLDoIwEAXQr2HbGW3qa8fCB2JiwKjQjUFTC4otKSh-vshOUXR2Mzk3uRngEABX0S2RUZFoFaXPnfd2rtdF5gwpLj0cYtd3OmNkLp32aQXCF-BPJxUYs87M3VBE_C-PX8b-mV8JBVvgrWzOGqBZswYtPebAZar39U9CW-3pQAI34iiMMORqqnNcFFk-stDCsi

Republic Of Trinidad & Tobago Ministry Of Education. (2017a). *Assessment Framework for the Secondary Entrance Assessment: 2019-2023*. Port of Spain, Trinidad and Tobago: Ministry of Education. Retrieved from http://moe.gov.tt/Education/Primary/SEA-2019-2023-Framework

Republic Of Trinidad & Tobago Ministry Of Education. (2017b). *Secondary Entrance Assessment 2017: Information Booklet*. Port of Spain, Trinidad and Tobago: Ministry of Education. Retrieved from http://moe.gov.tt/News/Publications

ReviseNow. (2018). English 2017. Retrieved June 20, 2018, from http://www.revisenow.net/index.php/english/year-2017

SCAA. (1997a). *English tests: mark schemes*. London, UK: School Curriculum and Assessment Authority.

SCAA. (1997b). *English writing booklet*. London, UK: School Curriculum and Assessment Authority.

Schools Week. (2017). Good riddance to secure-fit writing assessment. Retrieved July 25, 2018, from https://schoolsweek.co.uk/good-riddance-to-secure-fit-writing-assessment/

SEAB. (2015). *PSLE: English Language*. Retrieved from https://www.seab.gov.sg/home/examinations/psle/psle-syllabuses-examined-in-2019

SEAB. (2018a). National Examinations: PSLE. Retrieved June 18, 2018, from https://www.seab.gov.sg/pages/nationalExaminations/PSLE/general_information

SEAB. (2018b). *Primary School Leaving Examination: Instructions for candidates*. Retrieved from https://www.seab.gov.sg/pages/nationalExaminations/PSLE/general_information

Searchlight. (2014). CPEA holds teachers accountable – Head teacher. Retrieved June 18, 2018, from http://testwp04.newsmemory.com/searchlight/news/news/2014/06/20/cpea-holds-teachers-accountable-head-teacher/

Shorrocks-Taylor, D. (1999). *National testing: Past, present and future*. Leicester, UK: British Psychological Society.

Singapore Ministry of Education. (2016a). Changes to the PSLE scoring and

secondary one posting from 2021. Retrieved June 18, 2018, from https://www.moe.gov.sg/microsites/psle/

Singapore Ministry of Education. (2016b). PSLE Scoring: Streaming criteria. Retrieved July 20, 2018, from https://www.moe.gov.sg/microsites/psle/PSLE Scoring/psle-scoring.html

Singapore National Library Board. (2016). Primary School Leaving Examination. Retrieved June 18, 2018, from http://eresources.nlb.gov.sg/infopedia/articles/SIP_2016-07-08_114217.html

Smarter Balanced Assessment Consortium. (n.d.). Sample Items. Retrieved June 19, 2018, from http://sampleitems.smarterbalanced.org/BrowseItems?claims=ELA2&gradeLevels=56&subjects=ELA

Smarter Balanced Assessment Consortium. (2014). *English/Language Arts Practice Test Scoring Guide: Grade 5 Performance Task*. Retrieved from http://www.caaspp.org/ta-resources/practice-training.html

Smarter Balanced Assessment Consortium. (2017). *ELA practice test scoring guide: Grade 5*. Retrieved from http://www.caaspp.org/ta-resources/practice-training.html

SNSA. (n.d.-a). Questions and answers. Retrieved June 18, 2018, from https://standardisedassessment.gov.scot/questions-and-answers/

SNSA. (n.d.-b). What are the assessments like? Retrieved June 20, 2018, from https://standardisedassessment.gov.scot/more-about-assessments-and-reports

STA. (2015). *Interim teacher assessment frameworks at the end of key stage 2: September 2015*. Coventry, UK: Standards & Testing Agency. Retrieved from https://www.gov.uk/government/publications/interim-frameworks-for-teacher-assessment-at-the-end-of-key-stage-2

STA. (2016). *2017 teacher assessment external moderation: key stage 2 writing - For schools and local authorities*. Coventry, UK: Standards & Testing Agency. Retrieved from https://www.gov.uk/government/publications/teacher-assessment-moderation-requirements-for-key-stage-2

STA. (2017a). *2016 maladministration report: National curriculum assessments at key stages 1 and 2*. Coventry, UK: Standards & Testing Agency. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/655096/2016_Maladministration_report.pdf

STA. (2017b). *Teacher assessment frameworks at the end of key stage 2: For use in the 2017 to 2018 academic year*. Coventry, UK: Standards & Testing Agency. Retrieved from https://www.gov.uk/government/publications/teacher-assessment-frameworks-at-the-end-of-key-stage-2

STA. (2018). *Information for parents: 2018 national curriculum tests at the end of key stages 1 and 2*. Coventry, UK: Standards and Testing Agency. Retrieved from https://www.gov.uk/government/publications/key-stage-1-and-2-national-curriculum-tests-information-for-parents

TES. (2016). Ticking 198 boxes for new primary assessments means “days” of extra work. Retrieved July 25, 2018, from https://www.tes.com/news/ticking-198-

boxes-new-primary-assessments-means-days-extra-work

Testbase. (2018). Past SATs papers. Retrieved July 5, 2018, from https://www.testbase.co.uk/past-papers/

TGAT. (1988). *National Curriculum Task Group on Assessment and Testing*. London, UK: Department of Education and Science and the Welsh Office. Retrieved from http://www.educationengland.org.uk/documents/pdfs/1988-TGAT-report.pdf

The Observer. (2018). What is the top PLE school in Uganda? All 12,700 schools ranked. Retrieved July 20, 2018, from https://observer.ug/topics/56776-what-is-the-top-ple-school-in-uganda-all-12-700-sch...

The World Bank. (2017). Total population. Retrieved January 8, 2018, from https://data.worldbank.org/indicator/SP.POP.TOTL?year_high_desc=true

Thompson, G. (2013). NAPLAN, MySchool and Accountability: Teacher perceptions of the effects of testing. *The International Education Journal: Comparative Perspectives*, *12*, 62–84.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273–286. http://doi.org/10.1037/h0070288

UNEB. (2016). Primary Leaving Examinations. Retrieved June 18, 2018, from https://uneb.ac.ug/index.php/courses/primary7/?filter_action=course-breif

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & De Maeyer, S. (2019). Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. *Assessment in Education: Principles, Policy & Practice*, *26*, 59–74. http://doi.org/10.1080/0969594X.2016.1253542

Vincentian. (2017). The Caribbean Primary Exit Exam (CPEA). Retrieved June 5, 2018, from http://thevincentian.com/the-caribbean-primary-exit-exam-cpea-p13239-112.htm

Weigle, S. C. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press.

Whetton, C. (2009). A brief history of a testing time: national curriculum assessment in England 1989–2008. *Educational Research*, *51*, 137–159. http://doi.org/10.1080/00131880902891222

Whitehouse, C. (2012). *Testing the validity of judgements about geography essays using the Adaptive Comparative Judgement method*. Manchester, UK: Centre for Education Research and Policy. Retrieved from https://cerp.aqa.org.uk/research-library/testing-validity-judgements-using-adaptive-comparative-judgement-method

Whitelock, D. (2006). Electronic assessment: marking, monitoring and mediating learning. *International Journal of Learning Technology*, *2*, 264. http://doi.org/10.1504/IJLT.2006.010620

Published by:

ofqual

Earlsdon Park
53-55 Butts Road
Coventry
CV1 3BH

0300 303 3344
public.enquiries@ofqual.gov.uk
www.gov.uk/ofqual

March 2019

Ofqual/19/6492