BRILL

SCOTTISH EDUCATIONAL REVIEW (2025) 1–23

SER

brill.com/ser

Assessment Ambivalence: Teachers' Perceptions of National History Examinations in Scotland

Joseph Smith | ORCID: 0000-0002-4643-8388
Senior Lecturer in Education, Faculty of Social Sciences, University of Stirling, Stirling, UK
Joseph.smith@stir.ac.uk

Received: 7 February 2023 | Accepted: 27 May 2025 |
Published online: 5 September 2025

Abstract

Scotland's long tradition of high-stakes terminal examinations is coming under strain, as calls for diverse approaches to assessment increase in response to international criticism, and concerns about student wellbeing. This paper explores Scottish history teachers' views on the existing examination structure and finds a contradictory picture. Teachers are sharply critical of a system which they see as rigidly performative and argue strongly that it is of dubious validity. Despite this, Scottish history teachers continue to defend the system in terms of its reliability and manageability.

The paper proposes the concept of 'assessment ambivalence' to describe the cognitive coexistence of these paradoxical positions. The paper argues that teachers in Scotland are constrained by a lack of access to international research on historical understanding and assessment which leads them to accept an unsatisfactory status quo as a Leibnizian 'best of all possible worlds.' It concludes that access to these debates is a precondition for any improvements in the assessment regime.

Keywords

history – assessment – teacher perceptions

Scottish education has long relied on a system of high-stakes terminal assessments to award qualifications, and to determine student pathways, university

entry and school reputation (Wilson, 1975). Since 1997, most of these assessments comprise examinations devised, assessed and administered by the Scottish Qualifications Authority (SQA). In recent years, though, this system has come under scrutiny with academics and international reviewers highlighting its distorting influence on curriculum and pedagogy (Shapira & Priestley, 2018; OECD, 2021). This scrutiny intensified in the years during and immediately after the Covid-19 lockdowns when formal examinations were suspended. In 2021, the Scottish Government commissioned an Independent Review of Qualifications and Assessment chaired by Prof. Louise Hayward, which reported in 2023 (Scottish Government, 2023). The report recommended a new Scottish Diploma of Achievement which would aim to capture a more holistic sense of a child's achievement based on reduced use of high stakes terminal assessments. In 2024, the Scottish Government accepted many of the aspirations described in the report but adopted a more cautious approach to their implementation (Scottish Government, 2024).

In the context of these growing reservations, this paper asks about teachers' perception of high-stakes terminal examinations in one school subject, namely history. The data presented here are drawn from wide-ranging interviews with Scottish History teachers in which they were asked simply, 'What do you think of the examination system in History?'. It uses the familiar concepts of validity, reliability and manageability (Stobart, 2008) to explore these responses, and reveals a tension: on the one hand, teachers are frustrated that their pedagogical practices are distorted by the prescriptive SQA assessment regime, but on the other, they are grateful for this prescription as a set of guarantees which standardises their work. In the discussion section, the paper proposes the concept of 'assessment ambivalence' to explain this relationship, before questioning whether this ambivalent relationship can survive a recent dramatic drop in the pass rate for History Higher (SQA, 2024). The paper closes by arguing that the current approach to History examinations in Scotland has reached an inevitable end, and that international research about children's historical learning must inform a national debate about what we want children to do with history and how we might assess it.

Theoretical Context

This section explores the concepts of validity, reliability and manageability which are foundational to interpreting teachers' views on assessment. These concepts will be considered in general terms first, before they are applied to the question of assessments in history specifically.

Validity, Reliability and Manageability in overview

Assessment validity refers to the extent to which the results of the assessment accurately reflect the skills or knowledge being assessed. For instance, asking a child to write a poem is a valid test of poetry composition skills, as it directly assesses the intended skill. In contrast, reliability refers to the consistency of the judgement made in comparison to other students taking the same test. Tests on multiplication tables, for example, have a high degree of reliability – a child who scores 48/50 has clearly performed better than one who scores 4/50. There is often, however, tension between reliability and validity. A multiplication test *reliably* determines which child has memorised more times tables, but this says little about a child's overall mathematical ability. Similarly, tests on historical dates or European capitals may produce *reliable* data but offer limited insight into broader competencies. The opposite problem is observed in a poem which emphasises validity: while a child who writes a poem engages in a valid assessment of poetry composition, evaluating such work reliably can be challenging due to the subjective nature of poetry appreciation. It is difficult to confidently assert that one poem is better than another, and different assessors might reach different judgements.

To validity and reliability, Stobart (2008) added a third consideration for system level assessment: manageability. This is important because what might be feasible in terms of assessment depends on scale. While it is possible to devise tests that are both highly accurate and highly reliable (driving tests, for example) these are also resource intensive, expensive and time-consuming. A panel of examiners might be a valid and reliable assessment for the small number of students taking PhDs each year, but it cannot reasonably be scaled up to the tens of thousands of children in a school system. To conceptualise the irreconcilability of validity, reliability and manageability, Stobart offers the model of a one-handed clock which can only strike a balance between two of these considerations at once.

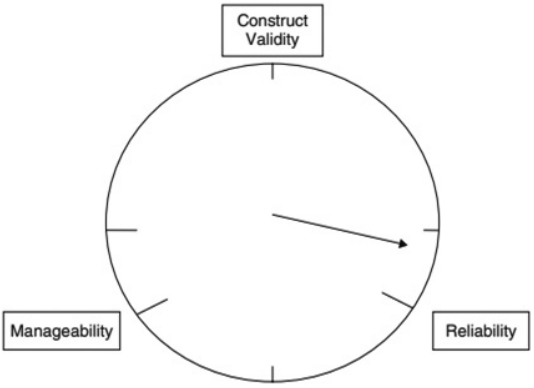

FIGURE 1 The one-handed clock (Stobart, 2008, p. 110)

Validity, reliability and manageability are especially important in highly centralised examination systems like the one in the senior years of Scottish high schools for two reasons. Firstly, the primary function of these examinations is to sift and sort – the grades awarded in National 5 examinations determine entry to study at Higher, and the grades awarded at Higher determine university entry. The system demands, therefore, that attainment data can be used to compare student performance – this is easier if all pupils sit the same examination according to the same assessment criteria. But examination results are not just used to record and classify student achievement, they also serve the same function in judging the 'performance' of the schools and teachers who taught these children. Both of these policy drivers privilege reliability over validity. If examination results are to be used to compare individuals and schools, it matters more that this data be 'robust' and 'objective' than that it be valid. However, these kinds of tests are wont to lead to distorted teaching and testing practices which undermine validity further (Smith & Fey, 2000).

The Search for Validity in Assessing Children's Historical Understanding

The difficulty of assessing historical understanding has blighted the subject ever since it was formalised as a discipline. In 1867, the nascent history component was removed from the Cambridge University Moral Sciences Tripos amid complaints that the subject represented little more than a memory test which rewarded the students who could cram more facts and dates than their colleagues (Slee, 1986). While these debates were later resolved in favour of an approach to historical assessment which demanded familiarity with historical sources, the question of how to develop a valid test which accurately reflects historical understanding (whatever that is) has never really gone away. That said, in a minority of national jurisdictions, questions of validity in assessing children's historical understanding remain unasked. Wineberg (2018) identifies the USA's SATs as a prime example of what happens when a system-level need for normal distribution curves and comparability of outcomes drives assessment practices. Wineberg argues that these tests select examination questions which ask about atomised 'factoids' purely on their ability to 'discriminate' pupils according to a normal distribution curve. He concludes, 'It's not sound historical judgment in the driver's seat, but the razzle-dazzle of the testing industry' (p. 17).

Fortunately, most testing regimes are less blasé about the importance of validity when assessing pupils' historical understanding. Seixas and Ercikan (2015, p. 1) identify three considerations in designing historical assessments: '(1) define models of cognition and learning in historical thinking, (2) design tasks and assessments targeting historical thinking, and (3) validate score meaning in those assessments.' The order of operation is crucial here: a clear understanding about what historical thinking is must precede any attempt to elicit or quantify it. Seixas (2017) identifies two dominant conceptions of history as a cognitive process: 'historical thinking' (which valorises proficiency in the disciplinary competences) and a 'historical consciousness' approach (which valorises an individual's ability to relate past, present and future). SQA assessment (SQA, 2021) is underpinned by the historical thinking approach which identifies 'subject skills' (p. 3) such as 'evaluating the origin, purpose, content and/or context of historical sources' (ibid) which pupils should develop. While such statements make clear SQA's position in relation to the first of Seixas and Ercikan 'three considerations', the second and third are more problematic: how to design tasks which access these cognitive models, and how to quantify these intellectually valid ways.

One way of measuring the validity of testing items in history has been to compare them to 'Think aloud protocols' (TAPS), in which respondents explain their thinking as they complete a given task (Smith, 2017). TAPS uncover what the respondent is thinking as they complete a task and so provide valid information about the cognitive process under scrutiny. However, since TAPS require a one-to-one assessor-student relationship, they do not provide the solutions in terms of manageability required at a system level. Research has supported the view that it is possible to devise written instruments in which written responses align closely with the kinds of thinking being assessed (Smith, et al., 2019), but only if the cognitive process being assessed is carefully and narrowly defined. Here, then, is a tension between the need for examinations to be transparent, predictable, and consistent from year-to-year, and the demands of validity. Reliability demands a somewhat mechanistic approach to marking which, in turn, encourages mechanistic responses and mechanistic teaching. Thus, an assessment which is designed to capture something complex – a pupils' relationship with the past – invites pupils to write according to a formula, or algorithm, that they know will hit the right notes on the mark scheme. Such concerns are not unique to SQA, nor are they new. More than thirty years ago, one school inspector in En-

gland complained that pupils were completing 'mechanical tasks rehearsing formulaic responses to snippets from sources' (Hamer, 1990, p. 24) cited in (Counsell, 2021).

In part, such responses are a product of a view of history education which reifies historical thinking, rather than historical consciousness. The fact that 'thinking' is a verb leads to an assumption that pupils can be called on 'to think' on demand, as they might be called on 'to dance' or 'to sing'. If thinking is considered a performance, we might expect performative responses. Outside the Anglophone world, an emphasis on historical consciousness has led to a more creative and holistic view of assessment. In addition to the familiar competencies demanded by a historical thinking model, Korber and Meyer-Hamme (2015, p. 93) add 'orientational competence' which 'defines the ability to relate information and insights about the past, as well as others' conclusions and judgments about the past, to one's own life (including one's society)'. In Quebec, Duquette (2015) used inductive analysis of pupils' responses to contemporary questions around poverty, immigration, and army enlistment to develop a four-level model for assessing pupils' historical consciousness. She argues that assessment instruments which focus on the way that pupils use historical information to respond to contemporary questions actually 'simplif[y] the process of evaluation since teachers do not need to evaluate every singly element of historical thinking individually' (p. 61). Similar research has been conducted in The Netherlands around using history to inform pupils' understanding of 'Enduring Human Issues' (Van Straaten, et al., 2018). In Sweden, national examinations invite students to compare how social phenomena we see today have changed over time (Rosenlund, 2021).

Context – Assessing History in Scotland's 'Senior Phase'

Scotland's curriculum divides pupils' schooling into two distinct 'phases': between the ages of 3 and 14 pupils study a 'Broad General Education' (BGE) which comprises curricula and assessments devised by the school, but from age 14 onwards pupils enter a Senior Phase in which they study for nationally recognised qualifications awarded by the Scottish Qualifications Authority (SQA), an executive non-departmental public body of the Scottish Government. This misalignment between a 3-14 phase which emphasises teacher assessment and school designed curricula, and a Senior Phase

which relies on a highly formalised examination system has been highlighted as a particular tension in the Scottish system (OECD, 2021; Education Scotland, 2024).

In History, SQA qualifications at Level 5 and above award 80% of the available marks for pupils' performance in two unseen examination papers covering Scottish, British and World History. Although these examination questions are unseen and undertaken in highly controlled examination settings, the format of the examination remain constant from year-to-year. Thus, pupils at Higher can expect to be asked 'How fully' an unseen source explains a given issue, with the source and the issue changing in each iteration of the test. This consistency of question structure is an important dimension of SQA's quality control process and is highlighted in curriculum documentation (SQA, 2021, p. 36; SQA, 2019, pp. 27-28).

A further guarantee of reliability for SQA is a transparent but rigid approach to awarding marks. In other words, teachers know what to expect of examiners and can therefore tailor their teaching accordingly. Figure 2 shows an example.

In addition to these written marking instructions, SQA offers in-person 'Understanding Standards' training for teachers which explains examiners' expectations in greater detail. More recently these sessions have been offered as live webinars which are also recorded and published online (SQA, 2021). However, as Stobart (2008) has pointed out, this entirely understandable emphasis on reliability and manageability might come at the expense of validity. For example, Figure 2 shows that there is a cap (6 marks) on the number of 'points of knowledge' that can be credited, meaning that a student who makes seven valid points rather than six, receives no additional recognition. This view that 'knowledge' in history comprises quantifiable atomised statements is intellectually insupportable and has been raised in the professional publications of Scottish history teachers (Philp, 2021; O'Hanlon, 2018). So far, these arguments have not received a wide audience.

The problematic rigidity of the SQA mark scheme in Figure 2 can be further highlighted by a comparison with the AS-Level Markscheme from AQA in England (notionally at the same level of challenge). Figure 3 shows an essay mark scheme which both assumes greater variability in candidate response and leaves more space for subjectivity and assessor judgement.

	Mark	3 marks	2 marks	1mark	0 marks
Historical context	**3**	Candidates establish at least two points of relevant background to the issue **and** identify key factors **and** connect these to the line of argument in response to the issue.	Candidates establish at least one point of relevant background to the issue **and** identify key factors **or** connect these to the line of argument.	Candidates establish at least one point of relevant background to the issue **or** identify key factors **or** a line of argument.	Candidates make one or two factual points but these are not relevant.
Conclusion	**3**	Candidates make a relative overall judgement between the different factors in relation to the issue and explain how this arises from their evaluation of the presented evidence.	Candidates make an overall judgement between the different factors in relation to the issue.	Candidates make a summary of points made.	Candidates make no overall judgement on the issue.
		6 marks			**0 marks**
Use of knowledge	**6**	**Up to a maximum of 6 marks**, award **1 mark** for each developed point of knowledge candidates use to support a factor or area of impact. Award knowledge and understanding marks where points are: • relevant to the issue in the question • developed (by providing additional detail, exemplification, reasons or evidence) • used to respond to the demands of the question (for example, explain, analyse)			Candidates use no evidence to support their conclusion.
Analysis	**6**	**Up to a maximum of 6 marks,** award **1 mark** for each comment candidates make which analyses the factors in terms of the question. Award **a maximum of 3 marks** where candidates make comments which address different aspects of individual factors.			Candidates provide a narrative response.
		4 marks			
Evaluation	**4**	Award **4 marks** where candidates connect their evaluative comments to build a line of argument that recognises the issue and takes account of counter-arguments or alternative interpretations.	Award **3 marks** where candidates connect their evaluative comments to build a line of argument that recognises the issue.	Award **2 marks** where candidates make isolated evaluative comments on different factors that recognise the issue.	Award **1 mark** where candidates make an isolated evaluative comment on an individual factor that recognises the issue.	Candidates make no relevant evaluative comments on factors.

FIGURE 2 Generic Mark Scheme for a 20-mark extended response at Higher Level (SQA, 2019, p. 05)

Generic Mark Scheme

L5: Answers will display a good understanding of the demands of the question. They will be well-organised and effectively communicated. There will be a range of clear and specific supporting information showing a good understanding of key features and issues, together with some conceptual awareness. The answer will be analytical in style with a range of direct comment leading to substantiated judgement. **21-25**

L4: Answers will show an understanding of the question and will supply a range of largely accurate information which will show an awareness of some of the key issues and features. The answer will be effectively organised and show adequate communication skills. There will be analytical comment in relation to the question and the answer will display some balance. However, there may be some generalisation and judgements will be limited and only partially substantiated. **16-20**

L3: The answer will show some understanding of the full demands of the question and the answer will be adequately organised. There will be appropriate information showing an understanding of some key features and/or issues but the answer may be limited in scope and/or contain inaccuracy and irrelevance. There will be some comment in relation to the question. **11-15**

L2: The answer will be descriptive or partial, showing some awareness of the question but a failure to grasp its full demands. There will be some attempt to convey material in an organised way although communication skills may be limited. There will be some appropriate information showing understanding of some key features and/or issues, but the answer may be very limited in scope and/or contain inaccuracy and irrelevance. There will be some, but limited, comment in relation to the question and statements will, for the most part, be unsupported and generalist. **6-10**

L1: The question has not been properly understood and the response shows limited organisational and communication skills. The information conveyed is irrelevant or extremely limited. There may be some unsupported, vague or generalist comment. **1-5**

Nothing worthy of credit. **0**

FIGURE 3 Generic Mark Scheme for a 25 mark extended response at AS-Level (AQA, 2019, p. 10)

Methods

This paper draws on interview data collected in the summer of 2018 as part of the 'Scottish History Teachers' Epistemologies Project' funded by the Carnegie Trust for the Universities of Scotland, which consisted of wide-ranging interviews (60-90 minutes) with 21 history teachers in Scotland. The research was granted ethical clearance by the University of Stirling in line with the guidelines laid down by BERA. Volunteers for interview were found through an open call on suitable Facebook pages and through the mailing list of the Scottish Association of Teachers of History (SATH). This convenience sampling approach raised the potential for selection bias: that members of these groups might be inherently more engaged in debates around history education and therefore not representative of the profession more widely. Nevertheless, it was clear in teacher responses that they perceived their answers to be a consensus view, accepted wisdom in their professional circles. These interviews discussed a range of topics, but all interviews included the question: 'What do you think of the examination system in History?' to elicit teachers' views.

In total, 21 teachers were interviewed, but quotations are drawn from just ten participants (six male, four female) who are referred to as pseudonyms throughout. While a broad consensus emerged across the cohort, not all participants are quoted for reasons of space. A decision was taken to quote 10 participants to demonstrate the breadth of feeling, while keeping the data comprehensible to the reader. In both the whole sample and the sample of those quoted, there was a skew towards less experienced teachers and an approximate 50/50 split in terms of sex.

Pseudonym	Sex	Years of teaching experience
Frank	M	2-5
Carol	F	>10
Lucy	F	2-5
Mary	F	<2
Chris	M	>10
Richard	M	2-5
Jack	M	2-5
Debbie	F	2-5
Henry	M	>10
Stephen	M	2-5

These interviews were coded using NVivo and analysed in three stages. First, open coding was adopted and categories such as 'formulaic answers' and 'inflexible marking' began to emerge. In the second stage, it became apparent that these codes could be organised according to Stobart's (2008) three dimensions of assessment – validity, reliability and manageability. Finally, data was organised under the headings 'criticisms of SQA' and 'defences of SQA.'

Results

This section is divided into two subsections:
- Participant's criticisms of SQA assessment practice
- Participant's defence of the same practices.

In general terms, participants were critical of the validity of assessments which they felt were formulaic and exercised a distorting influence on their practice, but they defended these invalid assessments by drawing on the concepts of reliability and manageability.

Participants' Criticisms of SQA History Assessments

Participants were acutely aware of the existence of a 'formula' for answering questions. The two exchanges below were from two different respondents but have the same complaint: that markers demand particular words are used in responses. Although interviewed separately, both respondents used, unprompted, the same word for SQA's approach: 'pedantic'.

Interviewer:	What makes it difficult?
Frank:	So, it's just really how pedantic it can be. Let's say, 'Evaluate the usefulness', to pick up marks for content, you've got to… use the word, if you don't say whether it's useful or not, you don't get the mark… they can understand the source very well, but they can't actually [get credit for this], because they're not using these little words. Well, not little words, but they're not using those specific words, they will lose marks for that.
Interviewer:	You said the marking was 'pedantic', what did you mean by pedantic?
Carol:	Okay. So for example, in the question that asks them to evaluate the usefulness, if a pupil does not write the words 'this source is not useful because…' or 'this source is useful because…' or 'this is source is less useful because…' at the beginning of every single point, they don't get the mark.

While the teachers in the examples above complained that a child would not be credited for relevant knowledge unless they used the correct 'little words', in the example below, the teachers have the opposite complaint: that a child who uses SQA approved words would get credit even if he demonstrates no knowledge beyond that given to him in the source.

Lucy:	they can quote from the source that's there but they will not get the mark in the exam unless they say 'this is useful because I know it is accurate' and I'm like 'why does that matter?'

INT:	Even if they don't know it's accurate? They don't need to know it's accurate to use those words do they?
Lucy:	Not necessarily. And it's just one of these things that I find really tedious, you're teaching them this thing and, like, when are they ever going to do that again as well? So what use is that?
Mary:	Because I think partly through our teaching and the pressure of exams and stuff we can become very much like we're just getting them to repeat things parrot fashion and actually there's no understanding there. So then they'll sometimes write stuff that doesn't really mean anything…
INT:	Does it get marks?
Mary:	It does [laughs] get marks, yeah. It absolutely gets marks, whereas something that's very good but doesn't include words like 'this is accurate information' at the end will not get a mark.
Chris:	[we] really drill them on the marking instructions and they will get a better outcome than… a really good historian who has their own written voice and doesn't play the rules by the game…, and so we're not training historians, we're training automatons.

The picture painted here is rather stark: of 'automatons' who are 'drilled' to play by the 'rules of the game' and who 'repeat things parrot fashion'. It is hard to imagine a pedagogical approach which jars more with SQA's claim that candidates develop skills of 'critically evaluating a variety of views' (SQA, 2019, p. 2). Teachers were clear that these mark schemes exercised a distorting performative effect over their practice. This was not just a matter of teachers supporting less able students in understanding their mark schemes, but a more troubling consequence where teachers were discouraging more able students from expressing themselves or their ideas. Consequently, the mark scheme is seen as rewarding students who lack historical knowledge, and penalising those with more knowledge who deploy it in unfamiliar ways.

Richard was clear on the pedagogical implications:

> [a student I was teaching] got four for evaluation and two for analysis. So I was talking to her and said 'listen, you're actually putting too much analysis in this section, let's go back and use it elsewhere'. That leads me

> on to the fact that I've got an academic background, done a bit of teaching, one of my criticisms of the Higher essays is the way they lay them out is not the way I would [do]. If I was teaching at university I would not teach a pupil to write this way, it's very formulaic [...] I get my pupils to colour code their essays. So green for knowledge, orange for analysis, yellow for analysis plus and you look at their essays and they are all green, yellow, orange, green, yellow, orange, tick, tick, tick, tick, tick.

Similarly, Jack recounted an anecdote which (though unverified) reads like a *reductio ad absurdum* of the SQA approach:

> One of my colleagues marks the Higher and she was saying...this one school wasn't writing in paragraphs. They were splitting it in individual sentences, leaving a gap. There's your knowledge. There's your analysis. There's your analysis plus. If you're preparing people to go to university that's problematic.

Teachers' Defence of SQA History

In the quote above, Richard and Jack create the sense that SQA had led students to produce not an essay, but a simulacrum of an essay; SQA responses were a specific idiosyncratic artefact created by the assessment process itself. Given the stark picture these teachers paint, we might expect their criticism to be unqualified, but this was not so. After describing the patchwork response quoted above, Jack went on,

> I think when I did my Higher it wasn't very structured like that and there was just a criteria of, "this is the marks" and the marker read it and then made a judgment, so it was very subjective... [but] the SQA's perspective on that is that there's no consistency there between markers so they've made it so formulaic to allow for consistency. You can essentially school a pupil who maybe doesn't understand the topic to write an essay that will get them marks.

This response shows considerable ambivalence – Jack is torn between the curriculum he studied (with more open questions and less 'consistency') and the current approach which rewards formulaic responses. Of course, what Jack calls 'consistency' is, more properly, reliability – the ability to cross compare between assessors. While Jack's age meant that he had had experience of two different approaches and was able to weigh the relative merits of each, younger teachers who had only experienced the current approach to examination were unable to think beyond it.

Debbie: There's an awful lot to remember, an awful lot of content, as well a formula for different questions.

INT: Formula of a particular question?

Debbie: So for instance, to evaluate the usefulness, we use 'J and the Five Ws', so that's a formula they remember in order to answer that question.

INT: So, National Five is hard, you think, but then also, formulaic.

Debbie: I think it has to be. It has to be. It's either that, or you say, right, take this question, interpret it whatever way you want.

Debbie's concern is that schools and children might 'take the question any way they want'. The assumption here is that there is a binary choice between the tick-box approach to marking pursued by SQA and a relativist anarchy where it would be impossible to judge the relative quality of responses.

A similar point was made in relation to the Higher in this lengthy extract from Henry,

Henry: We think Higher is very good. Any complaint, that's not the right word really, but any reservations we would have about Higher really fall on the way that marks have been allocated in essays. Now what that does is it makes essay writing as a skill mechanistic, and that's not necessarily a good thing, however the great strength of it is that it allows you to assure quality right across the country because if the SQA say 'this is how an introduction must be done and if you do it this way you will be given two marks', then they can assess that nationally because everybody in the country who does it the correct way gets the marks.

INT: Isn't there a concern about SQA deciding what 'the correct way' is?

Henry: There is a concern about that but the downside of not having a standard which is set centrally is that everybody goes off and does their own thing and nobody knows what the standard is, nobody knows how you would mark these in a summative assessment situation.

INT: Earlier, you talked about there should be holistic marking, what did you mean by that?

Henry: I don't think it should be you get two for an introduction, two for a conclusion, six for knowledge, four for evaluation, six for analysis.

INT: Why not?

Henry: Because it allows the less able pupils to get much closer to the higher able [sic] pupils than they should do. There's nothing more infuriating as a teacher when you read an excellent essay and it gets a 16 which in my opinion is a very good grade. I tell my pupils 'if you get 15 or above you're doing well'. But then you read another person's who doesn't really know what they're doing but picks up the marks and they get a 13, and I'm like there's no way there's two marks of a difference between these pupils. And then all it takes is for the good pupil to mess up a source question and the other person to do what they normally do, and they might get the same grade.

In this response, Henry draws an implicit distinction between reliability and validity. He accepts the reliability of the mark scheme as allowing a shared standard between schools, but he questions the validity of this. His argument is that a formulaic approach to assessment 'allows the less able pupils to get much closer to the higher able pupils than they should do' – in other words, it is not rewarding historical understanding appropriately.

Henry, Jack and Debbie share an ability to perspective-switch between the classroom and the educational system. Each of them agrees that the SQA approach distorts their practice, but each of them also argues that this is *necessary* from a system perspective to ensure reliability. These system level considerations were even more explicit in other responses. The contribution below introduces the third dimension of Stobart's one-handed clock: manageability. After some (admittedly heavy-handed) provocation from the interviewer, Stephen is critical of SQA assessment but defends it in terms of practicality and logistics.

Stephen: [at] university [level] when it comes to structure, you'd expect a structure for an answer…

INT: But you don't get two marks for that and two marks for that and two marks for that.

Stephen:	You don't, but how else? Because the other way you could do it is holistically and how are you supposed to mark holistically when you've got 150 markers marking 130 scripts each.
INT:	Lots of jurisdictions manage it
Stephen:	They do but at SQA level for history it's not, it's 'this is the answer, this is what we're looking for or something round about'.
INT:	But aren't SQA just making kids do weird things to make life easier for them?
Stephen:	You could say that but again it comes back to the argument that it's the most popular chosen subject, so what d'you do? You could look at what they do in philosophy [marking] where they can go away for a wee weekend and look at the small number of kids that they've got in Higher and mark it holistically or National 5 and do it that way, but when you've got so many kids because every kid wants to do it, there needs to be a structure to the marking.

This view differs from a critique solely on the grounds of validity versus reliability. Stephen wishes the assessment were otherwise, but thinks it is simply unmanageable to organise large scale assessments on this basis. This response shows perspective-switching: although frustrated with the demands of the examination from his perspective as a teacher, he can empathise with the position that the examination board finds itself in.

Discussion

Accardi (2023) refers to an 'access paradox' which binds Scottish History teachers who see success in national qualifications as the primary engine of social justice. These teachers want pupils to achieve well in their examinations but also

> 'recognis[e] that the status quo of a results-focused exam culture is problematic, but plac[e] blame on the system claiming that there is nothing they can do about it other than work within it, and this indicates what they think their role is in the classroom: to teach to the test.'
>
> p. 6

The data presented in this paper suggests something slightly different. Accardi is correct to identify a sense of powerlessness on the part of teachers, but teachers' relationship with SQA is more ambivalent than Accardi suggests. While they are acutely aware that the mark schemes are flawed, that they credit invalid knowledge and that they distort teaching practices, there is also sympathy – or rather empathy – for SQA. Teachers value the consistency, predictability and quality assurance offered by the mark scheme, even if they resent the implications of this on their classrooms. The data in this paper suggest a consensus view that SQA assessment is invalid but that this is a price worth paying for reliability. Put in the starkest possible terms, teachers were content to work within an assessment regime that they freely admitted was problematic, as long as it was equally problematic for everyone.

This ambivalence assumes something of a false binary between the dependability of SQA's intransigent marking practices, and a wild west free-for-all in which it is impossible to distinguish the quality of one response in relation to another. Teachers were freely critical of SQA practices, but they were not able to articulate how it could be otherwise. Jack spoke about his own experience of examinations as a pupil when 'the marker read it and then made a judgment' but even he was sympathetic to the SQA view that 'there's no consistency there between markers so they've made it so formulaic to allow for consistency.' Among younger teachers, there was even less ability to describe an alternative approach to assessment with comments such as 'it has to be like that' (Debbie) and 'how else are you going to do it' (Stephen). Unable to conceive of how things might be different, teachers are happy to delegate responsibility for student assessment to an arms-length body, albeit one about which they are uniformly critical. As I have written elsewhere, 'for many history teachers in Scotland, SQA remains not just the accountability mechanism by which their performance is judged, but also the only available conceptualisation of historical knowledge' (Smith, 2019 p. 18).

This inability to conceive of alternatives is even more stark when we see teachers making arguments from manageability. While we might present a defence of reliability as professional self-interest – teachers want transparency and comparability between schools – teachers also defended formulaic mark schemes because they made things easier for the examination authority at a system level. There is, on the face of it, no real reason for teachers to adopt this empathetic stance towards national examination bodies: manageability concerns are not educational, but financial and logistical. We can only speculate about the reasons for this surprising willingness to defend efficiency at the expense of validity. One explanation may lie in the size of the Scottish educational community. Scotland is a small country with a population of just 5.5 million and so the Scottish Qualifications Authority is not a distant bureaucracy, but

an organisation staffed with friends and colleagues. It is less surprising that teachers easily switch to the perspective of the examination authority when so many teachers act as examiners or know colleagues who do.

In the data we see 'assessment ambivalence': a mindset which emerges from the tensions between validity, reliability and manageability. This ambivalence can be summarised in the following train of thought which forms something of a consensus among the participant teachers,

> These examinations don't test what we would like them to (V), but at least they are the same for everyone (R) and, in any case, any test which did test what we wanted would be too complicated to administer (M).

Significant to this rationale is the switch in perspective that takes place midway: from the classroom-level to the system-level. This perspective-switch means that we are seeing something more complex than the familiar problem of 'teaching to the test'. These participant teachers are not just saying, 'I need to teach my pupils like this so that they succeed in the examination' (Accardi's 'access paradox' (2023)), they are also expressing sympathy for the very structures which shape these practices.

For many years, this ambivalence underpinned something of an unspoken agreement between teachers and SQA – teachers compromised the creative and expressive aspects of history in return for a predictable examination structure that allowed them to train pupils to succeed. Participant teachers might have been frustrated by, in Chris' words, 'training automatons' who 'play by the rules of the game' but this frustration was tempered by the knowledge that these students would be successful in their History qualifications. We might ask what would happen if this implicit pact between teachers and SQA were breached. Fortunately (or unfortunately), we do not need to ask this question in the hypothetical: shortly before this paper was submitted, the history examinations became a national news story in Scotland when the pass rate at Higher dropped to 65.7% from 78.7% the previous year (SQA, 2024). Although an official report by SQA blamed this drop on a particularly weak cohort of students (SQA, 2024), newspapers continue to publish accusations by teachers that the marking standard was artificially raised, disadvantaging a cohort of pupils (McEnaney, 2024; Seith, 2025). While it would be unwise to comment too much on an evolving story (McEnaney, 2025), it is possible to see this outrage as, in part, a howl of betrayal at a fracture in the symbiotic relationship between teachers and SQA, a relationship so long underpinned by the delicate balance of assessment ambivalence.

Ultimately, Scottish history teachers seem caught in a bind: they are critical of existing practices but unable to articulate how it might be different. Conse-

quently, among the dissatisfaction, there is resignation: 'if it could be otherwise, it would be otherwise'. There are echoes here of Leibniz's 'Best of All Possible Worlds' response to the problem of evil: Leibniz did not mean that our world was without considerable flaws but, that since this is the world that a loving God has chosen, the existence of better one was an impossibility. Key to disrupting this fatalism is the need to develop alternative discourses of history education which allow teachers to 'think against' the structures within which they work. These alternative discourses – drawn from a familiarity with international debates about what historical learning is, what it looks like, and how it can be assessed – will benefit both practising teachers and SQA itself. I have written elsewhere about the transformative power of these discourses in fomenting intellectually informed resistance in England (Smith, 2020; Smith, 2017b). The much-needed change to the system of History assessment in Scotland can only occur when these discourses reach a wider audience, and viable alternatives are proposed.

Implications and Conclusion

This paper has proposed the concept of 'assessment ambivalence' to describe the widespread view among Scottish history teachers that the examination is unsatisfactory but unimprovable. Assessment ambivalence gives rise to a form of pedagogical paralysis in which teachers adopt classroom approaches in which pupils are 'drilled' and 'repeat things parrot-fashion', leaving teachers disillusioned and unfulfilled. The relationship between uninspiring pedagogy and unsatisfactory assessments is a vicious circle, but one which Scottish teachers cannot see beyond. The SQA discourse is hegemonic in Scotland and, consequently, forms the horizons of Scottish history teachers' imaginations. It has been argued that this ambivalence and paralysis can only be disrupted through access to alternative conceptions of history education drawn from international debates, but how might this happen in practice?

At the time of writing, developments in Scottish education provide opportunities to break the cycle of rigid assessments and formulaic teaching. The Independent Review of Qualifications and Assessment (Hayward Report) recommended a shift away from high stakes terminal assessments (Scottish Government, 2023), and while the Government response to this was cautious, it accepted the broad case for reform (Scottish Government, 2024). Coinciding with this growing appetite for reforms to assessment, in September 2024 a new wide-ranging Curriculum Improvement Cycle was launched (Education Scotland, 2024). In its 'Case for Change', Education Scotland identified the need to address 'the disconnect between BGE and the Senior Phase' and suggested there needed to be 'greater clarity on what

knowledge learners should have' (Education Scotland, 2024, p. 25). Taken together, these reforms promise to address a long-standing problem in Scottish education – the misalignment between the 14-17 'Senior Phase' curriculum and the curriculum that children under 14 follow (OECD, 2021). To achieve this alignment, Education Scotland has concluded after consultation that a 'Know-Do-Understand approach should be used across the BGE and senior phase' (Education Scotland, 2024b, p. 32).

Taken together, these changes create an important opportunity to reinvent History education in Scotland away from the performative rigidity of existing assessment practices. There are many ways of thinking about school history (Smith, 2024) and the ongoing Curriculum Improvement Cycle allows a chance for an informed debate about the kinds of historical learning that it wants for its children and young people. Though traumatic the recent rupture in the relationship between History teachers and SQA triggered by the falling Higher pass rate (McEnaney, 2024), might create the space for all involved to consider more radical reforms to assessment than they might otherwise have entertained. This does not just mean a less rigid approach to assessment, but potentially a more fundamental reflection on what school history might be for.

Author Biography

Joseph Smith is a Senior Lecturer in History Education at the University of Stirling, UK. His research concerns the framing of history curricula, the identity of history teachers, and the intersections between these. He is a former high school history teacher and has published on both history and education for both general and academic audiences, most recently in the *Journal of Curriculum Studies* and *Journal of Education Policy*.

References

Accardi, C., 2023. *"It's rote learning, but it gets the mark": Exploring Scottish history teachers' (dis)engagements with social justice education.* Unpublished PhD Thesis: University of Strathclyde.

AQA, 2019. *AS History Mark Scheme.* [Online] Available at: https://filestore.aqa.org.uk/sample-papers-and-mark-schemes/2019/june/AQA-70412Q-W-MS-JUN19.PDF [Accessed 30 November 2021].

AQA, 2022. *AS History Mark Scheme* [Online] Available at: https://www.aqa.org.uk/subjects/history/as-level/history-7041/assessment-resources?secondaryResourceType=Mark+schemes [Accessed 27 June 2025].

Counsell, C., 2021. History. In: A. Sehgal Cuthbert & A. Standish, eds. *What should schools teach?.* London: UCL Press, pp. 154-173.

Duquette, C., 2015. Relating Historical Consciousness to historical thinking through assessment. In: K. Ercikan & P. Seixas, eds. *New Directions in Assessing Historical Thinking.* Abingdon: Routledge, pp. 51-63.

Education Scotland, 2024b. *Curriculum Improvement Cycle (CIC): Towards an Evolved Technical Framework,* Edinburgh: Education Scotland.

Education Scotland, 2024. *Curriculum Improvement Cycle: Background and Case for Change.* [Online] Available at: https://blogs.glowscotland.org.uk/glowblogs/public/cices/uploads/sites/10666/2024/11/13214842/CIC-A-CASE-FOR-CHANGE-PILOT-CURRICULUM-REVIEWS141124.pdf [Accessed 7 March 2025].

Education Scotland, 2024. *The Curriculum Improvement Cycle (CIC).* [Online] Available at: https://education.gov.scot/curriculum-for-excellence/about-curriculum-for-excellence/the-curriculum-improvement-cycle-cic/ [Accessed 7 March 2025].

Hamer, J., 1990. Ofsted and History in Schools. *The Historian,* Volume 53, pp. 24-25.

Harlen, W., 2005. Teachers' summative practices and assessment for learning—tensions and synergies. *The Curriculum Journal,* 16(2), pp. 207-233.

Hopfenbeck, T., Flórez Petour, M.-T. & Tolo, A., 2015. Balancing tensions in educational policy reforms: large-scale implementation of Assessment for Learning in Norway. *Assessment in Education: Principles, Policy & Practice,* 22(1), pp. 44-60.

Korber, A. & Meyer-Hamme, J., 2015. Historical Thinking, competencies, and their measurement. In: K. Ercikan & P. Seixas, eds. *New Directions in Assessing Historical Thinking.* Abingdon: Routledge, pp. 89-101.

McEnaney, J., 2024. Teachers hit out at SQA over 'unfair' exam marking. *The Herald*, 21 August .

McEnaney, J., 2025. SQA under pressure to publish 'lessons learned' from marking scandal. *The Herald*, 4 March.

Meadows, M. & Billington, L., 2005. *A Review of the Literature on Marking reliability,* Manchester: AQA.

Murchan, D., 2017. Introducing school-based assessment as part of junior cycle reform in Ireland: a bridge too far?. *Educational assessment, evaluation and accountability,* Volume 30, pp. 97-131.

OECD, 2021. *Scotland's Curriculum for Excellence: Into the Future,* Paris: OECD Publishing.

O'Hanlon, C., 2018. The Challenge of Performativity in the certificated history classroom. *SATH Yearbook,* Volume 26, pp. 73-78.

Philp, M., 2021. Recovering Education: History in the post-pandemic classroom. *SATH Yearbook,* Volume 27, pp. 51-60.

Priestley, M. & Biesta, G., 2013. *Reinventing the curriculum: New trends in curriculum policy and practice.* London: Bloomsbury.

Priestley, M. et al., 2020. *Rapid Review of National Qualification Experience 2020,* Edinburgh: Scottish Government.

Rosenlund, D., 2021. 'Continuity and change', large-scale assessment and equity: a study of gender-related differences regarding conceptual knowledge. *The Curriculum Journal,* 32(2), pp. 315-333.

Scottish Government, 2008. *Building the Curriculum 3: A Framework for Learning and Teaching.* [Online] Available at: https://education.gov.scot/media/0cvddrgh/btc3.pdf [Accessed 25 November 2021].

Scottish Government, 2021b. *Advisor to the Scottish Government on the reform of SQA and Education Scotland: remit.* [Online] Available at: https://www.gov.scot/publications/advisor-to-the-scottish-government-on-the-reform-of-sqa-and-education-scotland-remit/ [Accessed 25 November 2021].

Scottish Government, 2021. *OECD review backs school curriculum.* [Online] Available at: https://www.gov.scot/news/oecd-review-backs-school-curriculum/ [Accessed 25 November 2021].

Scottish Government, 2023. *It's our Future: Report of the Indpendent Review of Qualifications and Assessment,* Edinburgh: Scottish Government.

Scottish Government, 2024. *Scottish Government Response to the Final Report of the Independent Review of Qualifications and Assessment,* Edinburgh: Scottish Government.

Seith, E., 2025. History teachers give 'damning' evidence in row over SQA grading. *TES*, 22 January.

Seixas, P., 2017. Historical consciousness and historical thinking. In: M. Carretero, S. Berger & M. Grever, eds. *Palgrave handbook of research in historical culture and education.* London: Palgrave Macmillan, pp. 59-72.

Seixas, P. & Ercikan, K., 2015. Introduction: the new shape of history assessments. In: P. Seixas & K. Ercikan, eds. *New Directions in Assessing Historical Thinking.* Abingdon: Routledge, pp. 1-14.

Shapira, M. & Priestley, M., 2018. Narrowing the Curriculum? Contemporary trends in provision and attainment in the Scottish Curriculum. *Scottish Educational Review,* 50(1), pp. 75-107.

Slee, P., 1986. *Learning and a Liberal education: The study of modern history in the universities of Oxford, Cambridge and Manchester 1800-1914.* Manchester: Manchester University Press.

Smith, J., 2017b. Discursive Dancing: Traditionalism and Social Realism in the 2013 English History Curriculum Wars. *British Journal of Educational Studies,* 65(3), pp. 307-329.

Smith, J., 2019. Curriculum coherence and teachers' decision-making in Scottish high school history syllabi. *The Curriculum Journal*, 30(4), pp. 441-463. http://dx.doi.org/10.1080/09585176.2019.1647861

Smith, J., 2020. Community and Contestation: A Gramscian Case Study of Teacher Resistance. *Journal of Curriculum Studies,* 52(1), pp. 27-44.

Smith, J., 2024. History in the Curriculum. In: E. Rata, ed. *Research Handbook on Curriculum and Education.* Cheltenham: Elgar, pp. 354-367.

Smith, M., 2017. Cognitive validity: Can multiple-choice items tap historical thinking processes. *American Educational Research Journa,* Volume 54, pp. 1256-1287.

Smith, M., Breakstone, J. & Wineburg, S., 2019. History Assessments of Thinking: A Validity Study. *Cognition and Instruction,* 37(1), pp. 118-144.

Smith, M. & Fey, P., 2000. Validity and Accountability in High-Stakes Testing. *Journal of Teacher Education,* 51(5), pp. 334-344.

SQA, 2019. *2019 History Higher: British, European and world History. Finalised Marking Instructions.* [Online] Available at: https://www.sqa.org.uk/pastpapers/papers/instructions/2019/mi_NH_History_all_2019.pdf [Accessed 25 November 2021].

SQA, 2019. *Higher Course Specification.* [Online] Available at: https://www.sqa.org.uk/files_ccc/HigherCourseSpecHistory.pdf [Accessed 25 November 2021].

SQA, 2021. *Higher Marking Instructions*. Available at: https://www.sqa.org.uk/sqa/files_ccc/mi_NH_History_British-European-and-World-History_mi_2021.pdf [Accessed: 21 July 2025].

SQA, 2021. *National 5 Course Specification.* [Online] Available at: https://www.sqa.org.uk/sqa/files_ccc/n5-course-spec-history.pdf [Accessed 25 November 2021].

SQA, 2021. *Understanding Standards: History.* [Online] Available at: https://www.understandingstandards.org.uk/Subjects/History [Accessed 25 November 2021].

SQA, 2024. *Attainment Statistics 2024.* [Online] Available at: https://www.sqa.org.uk/sqa/105123.html [Accessed 6 March 2025].

SQA, 2024. *Higher History Review finds marking standards did not change – independent, external scrutiny confirms conclusion.* [Online] Available at: https://www.sqa.org.uk/sqa/112172.html [Accessed 7 March 2025].

Stobart, G., 2008. *Testing Times: The Uses and Abuses of Assessment.* Abingdon: Routledge.

Van Straaten, D., Wilschut, A. & Oostdam, R., 2018. Exploring pedagogical approaches for connecting the past, the present and the future in history teaching. *Historical Encounters: A journal of historical consciousness, historical cultures and history ,* 5(1), pp. 46-67.

Wilson, J., 1975. Examinations in Scotland 1960-75. *Research Intelligence,* 1(2), pp. 31-33.

Wineberg, S., 2018. *Why Learn History when it is already on your phone?.* Chicago: University of Chicago Press.