# Measurement and accountability in Scottish Education

In 2021, a briefing paper, written for the Scottish Greens by University of Stirling education researchers[1], explored the nature, scope and impact of policies designed to govern the education system through the collection and measurement of data. The paper explored the background to such approaches, as well as exploring their impacts – both through a review of literature and some small-scale empirical research with teachers and other education professionals. This paper further extends this work. It is based on focus group research with teachers and school leaders. It explores the kinds of measurement and assessment data collected, the usefulness of this, and the impact of collecting this data on teacher workload and working practices.

## Methodology

These results draw on comments made in two focus groups convened in June 2023 – one with school leaders, and a second with classroom practitioners. Focus groups were preferred to individual interviews because of the tendency for participants to speak more freely when they feel they are accompanied by others who understand their situation and because of the richer responses generate when participants 'flesh out' or exemplify the answers of their fellow participants.

Permission to conduct these focus groups was granted by the University of Stirling General Ethics Panel in May 2023. The research adheres to the British Educational Research Association code or research ethics. Participants were given information on the research project which made it clear that this research was funded by the Scottish Green Party. Participants were asked to give informed written consent for their words to be used in this report and were reminded of their right to withdraw their data. The composition of the focus groups and the anonymous codes used after quotations are given below. Participants reflected a range of rural and urban schools.

| Focus Group 1 – Senior Leaders | Focus Group 2 – Primary practitioners |
|---|---|
| High School Depute Head (SL1) | Primary Principal Teacher (PP1) |
| High School Depute Head (SL2) | Primary Teacher (PP2) |
| High School Head (SL3) | Primary Teacher (PP3) |
| High School Head (SL4) | Primary Teacher (PP4) |
| Primary Head (SL5) | |
| Primary Depute Head (SL6) | |

Focus Groups were conducted online using Microsoft Teams and were recorded and transcribed. The same prompt questions were asked in both focus groups:

1. What data on pupil progress and attainment does your school collect? Who asks for this data? How is this data used?
2. What is the impact of these processes on pupil experience?
3. What is the impact of these processes on teacher workload?

The senior leader focus group lasted 80 minutes, while the primary teacher focus group lasted 60. A grounded approach was taken to analysis data – the videos of focus groups were watched alongside the transcription and emergent themes were highlighted. These emergent themes (and associated participant quotations) are arranged under two subheadings in the next section of this report.

[1] Priestley, M. & Bradfield, K. (2020). *Educational governance through outcomes steering: 'reforms that deform'*. Scottish Green Party/ University of Stirling.

- Political demand for quantitative data on attainment, rather than progress
- Problematic Assessment practices

## Findings

### 1. Political/ system demand for quantitative data on attainment

Summative assessment data can be beneficial to teachers by providing information on pupil progress to guide their future learning. However, participants in both focus groups expressed a concern that much data on pupil attainment was collected to feed the demands of a system, rather than to support learners. Participants were clear on the distinction between these two ways of using data.

> I kind of feel though that a lot of the time the assessment, these summative assessments that are done seem to be purely for that tracking purpose. I don't really see it practically coming back into class. (Primary PT)

> I think at school level, what we're doing is we are gathering data to ensure quality provision for our children - we are there to be servants to our community. Unfortunately, data is gathered elsewhere for political reasons. (Primary Head)

In both primary and secondary schools, the norm was for data to be reported to local authorities on 3 or 4 occasions throughout the year. There was, however, a difference in how this was done – some local authorities ask schools to report their data in a format that suits the school, while others have a single authority-wide tracking system. In all local authorities, there was an emphasis on literacy and numeracy attainment in the Broad General Education phase (BGE), with some authorities also collecting data in Health and Wellbeing. In the senior phase, high schools were required to report on the attainment of their pupils according to key metrics.

Both classroom teachers and school leaders felt it was unjust that the local authority collected data on attainment rather than progress. That is, schools are required to report on the percentage of children were 'on track' to reach 'expected level' (First Level in P4, Second Level in P7 and Third Level in S2). Several participants in both focus groups described tracking systems in which children were colour-coded as Red, Amber or Green based on their likelihood of hitting the expected grade. Teachers had real concerns about the equity implications of this system (known by the unattractive acronym RAGging). One Primary Depute commented:

> It's not a measure of progress, it's a measure of achievement…For many of our children, they will go from P1 to P7 as a 'Red child' in the in the RAGging system - they are not 'on track' - and that can be for many, many reasons… School[s are] under pressure to increase the percentage of *achievement*, not to demonstrate progress for every child, but to increase achievement.

More worryingly, the Depute went on to explain how a system which prioritised attainment rather than progress created perverse incentives which led to limited resources being directed away from those children who found school most difficult.

> The Amber children are where the gains are. The system itself is set up to write children off from a really young age and to put the very little support we have onto 'Amber children' because you can you can affect your figures with those kids, but you can't affect your figures with the 'Red children'. I've got a huge number of examples from this year of children who have made massive progress - particularly in P7 - but they haven't 'made the level'. And really making the level from the start was always unfeasible. But the *progress* they made is huge. It's not registered anyway. It's not recognised anywhere and there's no reward.

It was clear that these practices were not restricted to a few schools and were echoed by a Principal Teacher in the Primary Teachers Focus Group:

> It's about stats, it's about raising the percentage of children in your school [achieving expected level]. On paper that looks like are on track… and we were 70% on track for writing and now we're 75%. But actually, the percentage that were at the bottom are still at the bottom and actually are not making progress year on year on year. So, I think actually quite a big problem is there because… as a teacher I'm actually more concerned about that child who is way behind and to try and move them forward. I think that child in P6 who is just under the border is probably gonna turn out fine as they move on. But it's a cynical way of looking at it. It's about raising your figures. It's not about helping the children that really need it.

Participants in both age phases were clear that tracking and monitoring student attainment on a regular basis could have benefits for individual learners, but they felt that the aggregated cohort-level data demanded by local authorities was both less useful and more influential. In particular, there were concerns about the influence of politicians and elected representatives who were seen as having an important oversight function, but also demanding additional unnecessary work. The following extract be a high school head captures this:

> In our authority, that data that is presented is used for the target setting for local authority and is presented the Educational Quality Assurance Panel four times a year. So the councillors themselves as well, the elected members are all over this information. Which means that there's incredible pressure on colleagues, because of course children are not all running through education as a sausage machine. But there is not an acceptance that some years you would be better. So there is a target of continuing improvement year on year.

A particular concern among school senior leaders was that too many different kinds of data existed in the system and there was not always confidence that politicians understood the data they were using. Another High School Head commented:

> One of the complications in the system that I see is that… the local government benchmarking toolkit that [elected councillors] use is a different data set than the data set that schools use. Which is a different data set again from the National Improvement Framework Toolkit… [But this] creates some confusion when talking about how schools are doing, because it very much depends which cohort you measure, when you measure them, and how what you include. I think the most obvious example is are you measuring National Fives or Level Fives? Are you measuring every leaver or only your S4 leavers, are you measuring the children that were in school in August or the children who are in school in December? Those all make a difference to the final figures.

Among high school leaders, this had two effects. One was that significant time was spent generating the kinds of data that local authorities requested – often in the knowledge that this data was not useful to the school or learners. The second was that a lack of precision around data in the system was influencing in school practices. One high school head gave the following example:

> The system is not fair because…you can do a First Aid qualification at level 6. It takes 5 hours in school but it carries the same kudos in terms of the league tables as Higher Physics. You know that your league table that will be published in The Times will reflect kids doing one day qualifications, giving it the same value as young people who are contributing to 160 hour courses and with an exam at the end. And that's where in some authorities where I have worked, you can see the manipulation of data to make sure that schools don't look bad in final outcomes. That's the bit I disagree with.

Participants agreed that these short courses were enormously valuable and appropriately levelled by SCQF. The problem, they felt, was an accountability system that did not distinguish between a five-hour Level 6 course and a 160-hour Level 6 course. For the other High School Head, this was enormously frustrating, but there was also concern that changes to ill-thought-out changes to accountability metrics might throw out the baby with the bathwater.

> There's a there's an error in the way that we are filtering our data, so it would be perfectly possible to filter our data so we could record all national qualifications separately. It's perfectly possible to add filter to include all SCQF qualifications over 60 hours or over 90 hours. But what we don't want to do is take away those valuable opportunities for learners, because… these are really valuable learning qualifications for some youngsters. And if we don't value them and don't record them, then we risk going back to the old 'everybody's got to do Highers' model. Actually, the plurality of what we offer children is really important.

## Problematic Assessment practices

*Curriculum for Excellence* (CfE)was intended to be a break from the more prescriptive 5-14 framework that had preceded it. In particular, a decision was taken to avoid a tick-box approach to assessment which might exercise a distorting effect on pupil experience. Each subject area of CfE includes a number of 'Experiences and Outcomes' (Es and Os) organised into First, Second, Third and Fourth Level. As their name implies, Es and Os outline two quite distinct things – *Experiences* (how children should engagement with specific content) and *Outcomes* (things that children should be able to do as a consequence of having learned). Often these two things may be conflated into a single 'E and O'. For example, SCN 1-02a ('I can explore examples of food chains and show an appreciation of how animals and plants depend on each other for food') contains an experience of substantive knowledge (children must be taught about food chains) and an expected outcome (be able to show an appreciation of this).

This dual nature of Es and Os means that they are inappropriate tools for assessment and, indeed, they were never intended to be used for this purpose[2]. In 2015, an OECD report stated,

> Too many teachers are unclear what should be assessed in relation to the Experiences and Outcomes, which blurs the connection between assessment and improvement. Beyond existing terms, current assessment arrangements do not provide sufficiently robust information, whether for system-level policy-making, or for local authorities, or for individual schools or across CfE domains for learners and their teachers.[3]

Shortly afterwards – in January 2016 – the Scottish Government announced the National Improvement Framework, which had two immediate effects. The first was the introduction of 'Benchmarks' at each level for assessment purposes, the second was the introduction of computer-based National Standardised Assessments for Scotland (SNSAs). The effect of this policy was felt more acutely in primary schools where there was a new emphasis in local authorities on tracking and monitoring pupils against the benchmarks, in order to ensure that pupils were 'on-track' to achieve the 'expected level' but the end of P4 or P7.

Following these policy decisions, Scotland has ended up with an educational accountability system which faces in two-directions at once. On the one hand, there is the original vision of CfE which aimed to avoid a prescriptive and tick-box approach to curriculum design and learner assessment, on the other there is the neoliberal governance model instantiated in the National Improvement Framework, which relies on hard numerical data to set targets and drive improvement. The problem, of course, is that CfE was not designed with such a governance model in mind. The 2017 Benchmarks – which are now used to measure progress – are simply rewordings of the original Experiences and Outcomes, but these Experiences and Outcomes were never intended to be used to measure progress at a system level.

---

[2] Early documentation demonstrated a sensitivity towards the dangers of assessment driving the curriculum. The 2007 overarching cover paper for the draft Es and Os was explicit that they 'are not designed as assessment criteria in their own right'.

[3] OECD (2015) 'Improving Schools in Scotland' Improving-Schools-in-Scotland-An-OECD-Perspective.pdf P.11

Education systems analysts write about 'input regulation' (what schools are told to do) and 'output regulation' (how schools are judged)[4]. Some school systems – England being an archetype – have both strong input and output regulation; schools are told exactly what to do and are measured against this. CfE was conceived as a high-trust model with both weak input regulation and weak outcome regulation, in which schools and teachers were trusted to know what was best for their learners. Since the National Improvement Framework, Scotland has moved towards a system with weak input regulation and strong output regulation. This tension is already exercising a distorting effect on schools.

However, this new emphasis on tracking and monitoring was flawed from the beginning because the Benchmarks were derived directly from Es and Os which were never intended as assessment tools. Four emerging issues related to this were discernible in Focus Groups.

1. The invention of subdivided levels to enable real-time tracking;
2. Concerns about inconsistency in moderation and assessment practices across Scotland,
3. A proliferation of third-party assessment tools and inappropriate assessment practices;

### *Subdividing Levels*

CfE Levels were always intended as normative statements about the attainment of students, with an expectation that the majority of students would meet certain levels at certain points in their schooling.

| CfE Level[5] | Stage |
|---|---|
| **Early** | The final two years of early learning and childcare before a child goes to school and P1, or later for some. |
| **First** | To the end of P4, but earlier or later for some. |
| **Second** | To the end of P7, but earlier or later for some. |
| **Third and Fourth** | S1 to S3, but earlier or later for some. The Fourth Level broadly equates to Scottish Credit and Qualifications Framework level 4.<br><br>The Fourth Level experiences and outcomes are intended to provide possibilities for choice and young people's programmes will not include all of the Fourth Level outcomes. |
| **Senior Phase** | S4 to S6, and college or other means of study. |

While this assessment framework would make it possible to report on individual pupil attainment at certain checkpoints (the end of P4, P7 and S3 are implied here), it is less useful as a tool for tracking cohort attainment over time. Since this kind of real-time tracking data is demanded by local authorities and political leaders, an unofficial system of 'sub-levels' has emerged in council areas. In all Primary schools for which we have data, CfE Levels were split into three so that expected

[4] For example, see: Leat, D., Livingston, K. & Priestley, M. (2013). Curriculum deregulation in England and Scotland - Different directions of travel? In: W. Kuiper & J. Berkvens (Eds.), *Balancing Curriculum Regulation and Freedom across Europe*, CIDREE Yearbook 2013. Enschede, the Netherlands: SLO.

[5] https://www.gov.scot/publications/achievement-curriculum-excellence-cfe-levels/pages/2/

attainment could be monitored throughout the school career (e.g., 'expected attainment' might be Level 1.1 in P1, 1.2 in P2 and 1.3 in P3). It is important to note that these sublevels have no existence in CfE and are of questionable validity – respondents explained how a single 'E and O' might be subdivided into 'progressive skills'. This dubious subdivision of CfE Levels was particularly pronounced in one school, which looked for progression on a term-by-term, rather than year-by-year basis. The result of this was a subdivided level which was subdivided again (e.g., expected attainment before the Christmas of P3 was 1.3 Bronze, by Easter it was 1.3 Silver, and so on).

Teachers had two main concerns about this practice: first, that it was of no benefit to children (particularly young children for whom it would be incomprehensible).

> I think a lot of these assessments that we're all doing, ultimately they're not very directly benefited, beneficial to learners. We don't sit down and go, 'So you're not on track and therefore…' That information isn't necessarily shared with parents either. It is very much in-house for a very kind of strategic management purpose and benefit. So we can pass all these percentages on… It's easier to have conversation[s] with the older ones to say, 'Right, you know what, I think you missed this in the maths assessment, so we'll do a bit of practice on that'. But for the younger ones, it's just something they need to do and it's much harder to have that conversation with them. I don't really think they benefit.

Teachers' second concern about sub-levels related to their accuracy.

> It's now if they are bronze, you know they're quite a bit behind. If they're 2.1 and silver, they need a bit of support and 'Just 2.1'. They're fully on track to meet 2.1 by the end of P5…. And obviously, when we got together as a [local authority] cluster, we realised all the schools were doing something so different.

Since these levels do not exist in any curriculum documentation, it is hardly surprising that teachers struggle to allocate them 'correctly'. In assessment terminology, the difference is between *validity* (the extent to which a test measures what it is trying to measure) and *reliability* (the extent to which marks can be relied as fair). Teachers were keenly aware of the challenge of allocating reliable marks but are not challenging the concept of these minuscule sub-levels at the level of *validity*.

### *Concerns about moderation*

The previous participant's response draws on a wider concern within the focus groups that there was limited comparability of standards across Scotland. A particular source of debate concerned the 'amount' or 'percentage' of the benchmarks that children needed to 'complete' to be awarded a level. Such a view considers each of the Benchmarks as an outcome that can be 'completed' by a child and suggests a fatal misunderstanding of what the Es and Os and Benchmarks are for. It is wrong-headed in at least two ways. First, not all Benchmarks are equally 'large', meaning that 'ticking off' a benchmark and calculating a percentage makes little sense. Secondly, the Benchmarks are not things that a child demonstrates and then moves on from. Consider the following,

| | |
|---|---|
| ***I can convey information, describe events, explain processes or combine ideas in different ways.*** <br> ***LIT 2-28a*** | ***When writing to convey information, describe events, explain processes or combine ideas in different ways:*** <br>• ***Uses appropriate style and format to convey information applying key features of the chosen genre.*** <br>• ***Includes relevant ideas, knowledge and information.*** <br>• ***Organises and presents information in a logical way.*** <br>• ***Uses tone and vocabulary appropriate to purpose.*** |

The skills and competencies described here cannot be understood as a binary (child can/ cannot) because they refer to things that even the most accomplished authors do. They describe not a benchmark to be 'hit' but Age 11, but a constellation of skills in written communication. Nevertheless, the accountability mechanisms in the system demand that a child's attainment level

can be calculated in some way and teacher's concerns continued to be about the reliability of that judgment rather than its validity. As one Primary head stated,

> Moderation practises across Scotland are all very different. I would say even within our Associated School Group. So that the primaries that that move on to the same secondary and we're all measuring in a slightly different way. And I understand why the Scottish Government and the local authority won't say about the benchmarks, about what percentage you need to have completed before you have been successful at a level.

However, the same respondent appreciated that a central government pronouncement about the potential for perverse outcomes if government were to be more prescriptive about what it means for a child to 'achieve a level'.

> I think a lot of the moderation practises across Scotland needs to be looked at again. Having said that, I also understand [if the government says that] it's 50% or 70% of the benchmarks that we need, then that's what we'll do. We'll all just home in on those. So I fully understand why the government and the local authority won't give us guidance around that and it's up to the school to make that decision… But I do think because of that we are then judged on our results when they're not actually accurate.

### *Proliferation of assessment tools and inappropriate assessment practices*

The difficulty that schools encountered in awarding accurate levels has led to a proliferation in the number of third-party assessment tools used in schools. The Primary Focus Group comprised five participants who between them identified 15 different assessment tools in use in their schools[6]. Teachers were clear that schools and local authorities sometimes favoured assessment tools because of their ability to generate a robust calculation of a child's attainment, even if these were not necessarily the most pedagogically appropriate. Teachers in focus groups explained that there has been an increase in the use of pencil and paper tests conducted in exam conditions.

> More recently last kind of two years we've been using more standardised assessments, so PUMA and PITA they're quite tricky in terms of differentiation because there is an expectation that if you have the primary five class, all learners set the primary 5 paper… My stage partner [and I] had raised the concern with her head teacher to say, you know, that's not gonna work for our learners at all. It's so differentiated our class. It's just, there is no point in giving a primary 5 paper to child who still working at early level. So we raised that concern as you know, she said obviously, you do it as you see necessary for your learners. So we did that and obviously it's above a learning curve again that has been a nightmare because then we've had different papers across the P5 cohort [to set and mark].

Another Primary Teacher said something similar,

> It depends on who your management is in the school at the time whether you can actually say, 'I know for a fact that child will fail during that test because they've not got the skills to do it. And you know I've identified them as my professional judgement'. Sometimes you have the battle or just give the child a test anyway: a test to fail. Lovely!

When asked specifically about the SNSAs which were introduced as part of the National Improvement Framework, participants tended to see these irrelevant to their practice. One Primary Depute stated,

> SNSAs don't add it into us at all and they don't provide us with anything. So I would remove [them]and I would invest significantly in some form of national moderation programmes that might be able to.

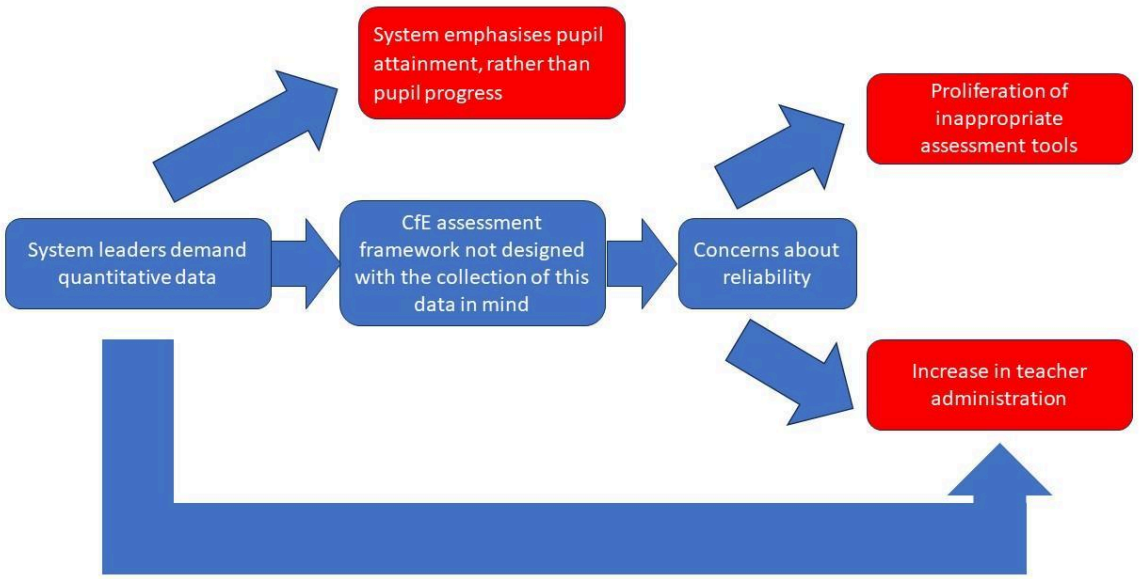
[6] GL Assessments, SEAL Numeracy, Suffolk Spelling, Highland Literacy, PITA tests, PUMA Tests, Ros Wilson Grids, Single-word spelling tests, Skills Development Scotland meta-skills progression framework, Shanarri Wheels, PIRA tests, Roots through writing, Diagnostic Numeracy assessments, Read Write Inc, Scholastic PM Benchmarks

## Conclusion

Underlying all these problems is a disconnect between the aims and purposes of CfE and the neoliberal governance models that have been imposed on top of it. When asked about the value of termly tracking and monitoring in a primary setting, A PT said:

> It's for the council. It's for ultimately, it's for the Council to know what their stats are and then it is for the management… I still think that the summative assessments ultimately are not still providing an overall benefit for a child in the holistic sense. And then [there's the] fact that we don't track other stuff. We don't track wellbeing properly. We don't track their wider achievements properly. We don't track their critical analysis, thinking and social studies and subjects like that, things that really matter in these little people that we're trying to develop as adults of the future.

As highlighted in the previous briefing paper (Priestley & Bradfield, 2021), the issue lies in a tension between, on the one hand, policies designed to enhance the quality of education through the design and support of inputs (e.g., well designed classroom curriculum and effective pedagogy) and, on the other hand, policies designed to monitor the performance of an education system through measurement. While it is necessary to collect meaningful data about the quality of the system, there is ample evidence (in Scotland[7] as exemplified in this paper and further afield) that crude use of data to achieve narrow accountability goals has a major effect of generating perverse incentives. This will over time result in the development of performative cultures in schools, as stakeholders such as teachers and school leaders seek to generate the right sort of data. The diagram below illustrates some of the processes that occur.




[7] Also see: Shapira, M., Priestley, M., Peace-Hughes, T., Barnett, C. & Ritchie, M. (2023*). Choice, Attainment and Positive Destinations: Exploring the impact of curriculum policy change on young people*. University of Stirling/Nuffield Foundation. This research surfaced some especially egregious examples of performativity, including the abolition of under-performing subjects in some secondary schools, regardless of their desirability as part of a broad and balanced curriculum).

Moreover, as schools jump to the data demands of the system, they potentially losing sight of the educational purposes that should drive practice, with negative consequences for teachers and ultimately for children and young people. We see a situation where schools and teachers operate to fulfil arbitrary system demands, rather than the inverse – a system that exists to support and facilitate good educational practice in schools and classrooms.