Secondary Phase Summative Assessment Principles

At Dixons, the most important assessment data in our academies comes from formative assessment. Every day, we use specific and repetitive minute-by-minute formative assessment and other leading indicators to help shape students' learning.

However, for our Trust and academy leaders, analysing summative data, even though it is lagged, can also help to inform interventions and dynamic resource allocation both in the moment and over time.

Curriculum alignment

- In each EBacc subject, there is an agreed minimum body of knowledge (substantive and procedural) that all academies are expected to cover by the end of Year 7, 8 and 9. How that knowledge is sequenced and taught over each year is up to the individual academy. In any given year, each academy can also extend the body of knowledge if they wish to.
- Leaders are also committed to aligning KS4 exam specifications and long term plans in the EBacc. Once again, each academy has the autonomy to sequence and teach knowledge as they see fit.
- A department with the confidence of the Executive, will be allowed to innovate (inc. teach a different specification); however, if their results do not match, or exceed, that of the best departments across our Trust, they will be expected to align to the direction set by the cross-cutting team champion.
- If several academies decide to align further, other academies must still be afforded the right to only align to the minimum body of knowledge and / or to innovate. However, all academies must continue to engage fully in cross-cutting teams.

Summative collection cycle

- Summative assessments need to be far enough apart that students have the chance to meaningfully improve. Also, on the large domains of content which most summative assessments sample, students will not make particularly rapid improvements.
- As each academy has the autonomy to sequence and teach knowledge as they see fit, we can be confident that the same content will be taught in each academy by the end of each year, but not by the end of any given week or term.
- Therefore, once again, if summative assessments are used too frequently, there are risks:
    - students and teachers get demoralised because hard work in class is not showing up as improvement
    - students and teachers start to focus on short-term tactics which will lead to improvement on the summative assessment (but not lead to real improvement in learning)
    - teacher workload increases alongside opportunity cost more broadly, since testing windows inevitably eat into teaching time and impose discontinuities into the flow of the school year
- It is because of this that our Trust only sets common assessments once a year (towards the end of Cycle 3) and we don't expect our academies to set more than one mid-year summative assessment (either at the end of Cycle 1 or Cycle 2).

Assessment scope

- Up until the end of Year 10, common assessments set by our Trust are cumulative[1] (rather than unit[2] or global[3]):
    - Cumulative assessments take advantage of the spacing effect: if you have already studied something, studying it again after a delay can produce a huge amount of learning
    - Knowing there will be a cumulative summative assessment changes the way most students study (for the better). Research suggests that simply telling students that there will be a cumulative assessment may enhance their learning
    - Students often underestimate the value of repeated studying and they do not like cumulative assessments for the very reason they ought to be used: preparing for them requires more time and energy devoted to understanding and remembering content
    - Cumulative assessments are a desirable difficulty: they enhance learning but students do not like them
- Once students have covered enough of the curriculum, common assessments can be full past papers i.e. global in scope. This shift in scope is unlikely to happen until the end of Year 10; it is curriculum-driven and decided by leaders on cross-cutting teams.
- Global assessments may be useful in reassuring teachers about their predictions; however, sitting full papers too early is unlikely to enhance learning and may cause greater anxiety amongst students.

[1] What has been studied thus far (since the start of Year 7)

[2] What has just been studied (since the previous assessment)

[3] Giving a full past paper irrespective of whether all elements have been studied thus far

Standardisation

- Perhaps the most important concept in assessment is validity. Daniel Koretz, Professor of Assessment at Harvard University, says that 'validity is the single most important criterion for evaluating achievement testing'. It is the central concept in assessment.
- The actual result on an assessment does not matter. What matters are the inferences that we can make from that result. We need to be sure that the assessments we use are capable of supporting significant inferences. The process of doing this is called validation.
- One difficult aspect concerning the validity of an assessment involves sampling. Most assessments designed to produce a summative inference do not directly test the entire domain. So, when thinking about the validity of a summative inference, we nearly always need to consider what the domain is that we are trying to measure and how the assessment has sampled from that domain.
- A second vital assessment concept is reliability: the consistency of assessment. One factor that can have a big impact on reliability is agreement between markers. Therefore, internal and external (where possible) moderation is carried out across our Trust.
- Reliability is particularly important to consider when it comes to measuring progress. When we measure progress, we are often looking at the difference in performance between one test and the next. As such, we have two sets of potential measurement error to deal with: the error on the first test and the error on the second. Understanding reliability helps us to understand whether students really have made progress or not. As a crude rule of thumb, we consider individual student percentile rank improvement using a +/- 5% range, but class / academy averages with more precision.
- Where possible, our common assessments are designed externally in order to make sure that our assessments are more valid and reliable. For example, externally designed assessments are less familiar or less predictable for teachers. If some teachers see the assessment in advance, they might distort the curriculum (or advice regarding revision topics) in a manner that improves assessment performance, but not the wider student knowledge domain.
- Our common assessments (Y7-11) challenge students and provide an experience that matches that of final GCSE examinations. Academies are expected to ensure that assessments are undertaken in standardised conditions where students and teachers have standardised perceptions of the importance of the assessment.
- For leaders to make valid inferences across classrooms, or across academies, they need to be clear that they understand how the stakes are being framed for *all* students taking the test, even those who are not in their own academy.

Approximation and grading

Trust and academy leaders

- At leadership level, we use percentile ranking as our means of comparing students and to track a student's progress over time and whether it diverges between subjects.
- Once students start to sit global assessments (at the end of Year 10 and in Year 11), we use age-independent grades (1-9) which provide a more accurate indication of a student's likely performance in their final GCSE exams.
- Assessment data is fundamentally about prioritisation and is used to inform leaders' decisions around where to invest resources. Class teachers use ranking data to support their intervention planning which is focused on critical knowledge gaps and misconceptions so that all students achieve at least a grade 5 or VA of +1 by the end of Year 11, whichever represents stronger performance.

Students and their families

- In Years 7, 8, 9 and 10 (up to end of Cycle 2), we simply record and report raw scores from assessments as they are (with academy averages to provide benchmarks) as this is clear fact and judgement free.
- We only start to report GCSE grades to students and their families once students sit full past papers that are global in scope. This is because GCSE grades are intended to be used to judge a student's understanding of a full specification.
- At the end of the year (only), for all years, we use colours to recognise those students with high attainment (purple), or who have made large shifts from their baseline (purple), as well as to prioritise students for academy-leader-driven intervention (red).

Explanatory notes

Formative and summative assessment

Assessment is a vital part of education. When assessment is done badly, it can get in the way of good teaching, and even distort it. But done well, assessment can give students better information about their performance and teachers better information about whether their methods are working.

At Dixons, the most important assessment 'data' in our academies comes from formative assessment. Every day, we use specific and repetitive minute-by-minute formative assessment to ensure that both the teacher and student understand three key things:

1. Where the learner is right now
2. Where the learner is going
3. How the learner can get there

However, we do not grade or even record this vital 'data'— at least not in any prescribed way. After all, as Daisy Christodoulou stated in her book 'Making Good Progress'[4], grading formative assessment would be like a marathon runner measuring their weight-training in hours rather than kilograms. At best, pointless, and, at worst, grossly misleading.

At Dixons, formative assessment is integrated and ongoing – it aims to be responsive not reportive. If we do record marks, these should not be converted to grades. When converting to grades, you are asserting that the difficulty of the two assessments is the same and that you are trying to derive a shared meaning. Also, the aim of formative assessments is to set questions that are closely tied to what is being studied.

Summative assessment serves a different purpose and is used to sum up students' learning at the end of a period – for example, at the end of a cycle or the school year. As it sums up what the students have achieved, it looks back and checks for the retention and application of a broader knowledge base from the subject domain. This is more like a marathon runner timing their practice races which is why we do believe it's useful to record this data.

Approximation and grading

1. Trust and academy leaders

At leadership level, percentile ranking is our means of approximating where a student resides along the Trust distribution. This can be used to track a student's position on the distribution over time and whether it diverges between subjects. Of course, individual student ranks can be very 'noisy', but this becomes increasingly attenuated as data gets aggregated across multiple students and academies. As such, approximating a class, year or academy's average position along the distribution can help to inform academy and Trust leaders' decisions around how to allocate resources and focus interventions.

However, we can only perform this approximation if our summative assessments are as valid and reliable as possible. This is why we have committed to Trust-wide end of year common summative assessments. The other obvious benefit of this approach is that it enables like-for-like comparisons between all of our students, regardless of which academy they attend. However, this comparability is dependent on us ensuring consistent test conditions as well as aligning our curriculum across all participating academies.[5]

At Dixons, summative data is an approximation of a student's position along the Trust distribution. In order to arrive at this approximation, we leverage consistency, scale and technology.

If all students across our Trust take the same blind assessment, at the same time, under the same conditions, having covered the same content, we have a sample of >1,000 students per assessment (e.g. for Year 9 geography). We can then capture the raw marks for each of these students to provide us with a Trust distribution curve.

As asserted above, the purpose of this summative data is to inform school and Trust leaders' decisions around where to invest the scarce resources at their disposal. In other words, this data is fundamentally about prioritisation, which is a distinctly relativist exercise. As such, so long as we can understand how students, classes and schools are performing relative to each other, it is not necessarily that important for us to know their exact positions along the national curve.

The most useful summative assessment data we have at our disposal is a student's Trust percentile rank. Therefore, by reporting relative progress and percentile rank, subject leaders and teachers can determine which students to intervene with post-assessment. As such, the focus of assessment is to determine relative progress and identify gaps in subject knowledge so that gaps are closed.

The most important assessment data in our academies comes from formative assessment. This data doesn't need to reside in any systems.[6] In fact, a formative assessment designed for responsive teaching needs to stay in the classroom. Leaders should understand that once teachers know they're seeing data too, the test's purpose is changed, and that can skew behaviours. Or, to put it another way, if leaders want teachers to be using formative assessments to spot gaps in learning, they shouldn't incentivise them to set easier

[4] Daisy Christodoulou (2017). Making Good Progress? The future of assessment for learning

[5] Rich Davies (2019). Making the Grade. https://medium.com/@richardrhysdavies/making-the-grade-4b57c2284142

tests where everyone gets 100% just to impress those who oversee their work. There are caveats – there is value in collecting some metadata on formative assessment use (e.g. % of students logged in to a system over the past week), providing the data collected doesn't distort behaviour.[6]

2. Students and their families

The topic of grading has been under some scrutiny, with some excellent blogs by Becky Allen and Matthew Benyohai adding to this older, but no less relevant, piece by Alfie Kohn on the potential folly of grading.

At Dixons, we find little to disagree with in these blogs. Choosing how to give attainment feedback to students and their families is a minefield, but we know that choosing NOT to give clear, interpretable (i.e. often norm-referenced) feedback on how a student is doing is not a neutral position to take.[7]

Every teacher has to give some feedback on attainment and there is no risk-free or value-neutral approach to doing it. Knowing a student's ability or attainment cannot help you predict how they will respond. What matters is how the grading information:

- changes students' beliefs about their attainment
- changes their beliefs about their ability to learn and get better
- changes their desire to keep playing the competition of trying to be the best, or maintain their position, or avoid the bottom rung[8]

In Years 7, 8, 9 and 10 (up to end of Cycle 2), we give students clear cohort referenced feedback on how they are doing. We simply record and report raw scores from assessments as they are. As scores only make sense as indicators of standards in comparison to something, we also share cohort averages.

This approach does come with risks, but also has the potential to be a powerful motivator. Given the strength of our culture and social norms at Dixons, we can be confident that our students will respond positively.

We only start to report GCSE grades to students and their families once students sit full past papers that are global in scope. This is because GCSE grades are intended to be used to judge a student's understanding of a full specification. This shift in scope usually happens at the end of Year 10; it is curriculum-driven and decided by leaders on cross-cutting teams.

The amount of knowledge studied in Years 7, 8, 9 and 10 (up to end of Cycle 2), relative to a specification and the wider domain, is neither broad nor deep enough to extrapolate and benchmark performance against GCSE grades. Attempts to generate grade boundaries and imagined trajectories (worked back from GCSE grades) before the end of Year 10 are falsely linear and fictional. When you give students a grade it means something to them. Saying they are on a grade 6 in Year 9, but will improve, is a fabrication. Predicting a grade 6 for a student when they are just as likely to achieve a 5 or 7 is worrying. Calling a question, a 'grade 9' question, means that students who don't get all the marks don't think they can get a grade 9. This is unfair.[9]

However, at the end of the year (only), even for those year groups for whom we don't share GCSE grades, leaders do use the derived approximations of our students' current performance relative to their Trust peers to recognise those students who perform exceptionally well on an assessment and identify those who should be prioritised for academy-leader-driven intervention. As a crude rule of thumb, we consider individual student changes using a +/- 5% on percentile rank, but class / academy averages with more precision. Therefore, we can be confident that a percentile rank of ≤30% is likely to be below the national expectation and a rank of ≥ 80% is likely to represent high attainment. The goal for all students is to achieve at least a grade 5 at the end of Year 11:

prioritise for leader intervention		**recognise for high attainment**
percentile rank ≤ 30%	Percentile rank 31% to 79%	percentile rank ≥ 80%

Once GCSE grades are applied (usually from Cycle 3 in Year 10) the following applies:

prioritise for leader intervention		**recognise for high attainment**
Grade ≤ 3	Grade 4 to 6	Grade ≥ 7

For mid-year academy-based assessments, we don't report these colours; however, by comparing a student's academy percentile rank on a mid-year assessment to their academy rank on-entry, it is still possible for leaders to identify those students who have made large shifts and deserve special recognition (e.g. top 'mountain climbers').

[6] Joshua Perry (2020). What kind of analytics does a school or MAT leader need? http://www.renlearn.co.uk/renaissance-blog/analytics-school-mat-leader

[7] Becky Allen (2019). Writing the rules of the grading game (part III). https://rebeccaallen.co.uk/2019/04/27/grading-game-part-iii/

[8] Becky Allen (2019). Writing the rules of the grading game (part III). https://rebeccaallen.co.uk/2019/04/25/grading-game-part-ii/

[9] Matthew Benyohai (2019). Banning GCSE grades before Year 11. https://medium.com/@mrbenyohai/banning-gcse-grades-before-year-11-8737b40180a

Progress

Measuring progress – the change in attainment between two points in time – is a recent trend, but is very difficult to do reliably, and progress trackers, like flightpaths, fundamentally don't work for many subjects.

When we use tests to measure relative progress, we often look to see whether a student has moved up (good) or down (bad) the bell curve. If a student achieves a higher position on the bell curve on an end of year test than they did on a test taken at the start of the year, it looks like they've made good progress and learnt more than similar students over the course of the year. However, the student's test scores are a noisy measure of what the student knew at the start and end of the year. Therefore, neither test is reliable enough to say if this individual student's progress is actually better or worse than should be expected, given their starting point.

The challenge in measuring progress between two testing points is considerable because, whether we like it or not, the progress of individual students is slow compared to the variability of achievement within the age cohort. This means that a school will typically find that only a minority of their pupils record a test score growth statistically significantly different from zero. [10]

For further disavowal of 'progress' as a meaningful numerical construct, take a look at Tom Sherrington's blog post on the matter and this blog by Matthew Benyohai.

Also, measuring student progress is worse than irrelevant because intervening on progress data is frequently unjust and disadvantages those who have historically struggled at school. Suppose you find two students who get 47% in your end of Year 7 history test. It isn't a great score and suggests they haven't learnt many parts of the year's curriculum sufficiently well. Will you intervene to give either of them support? The response in many secondary schools nowadays would be to interpret the 47% in relation to their KS2 data. For the student who achieved good scaled scores at age 11 of around 107, the 47% suggests they are not on track to achieve predicted GCSE results and so will make a negative contribution to Progress 8. They are therefore marked down for intervention support. The other student left primary school with scaled scores around 94, so despite their poor historical knowledge at the end of Year 7, they are still on track to achieve their own predicted GCSE results. No intervention necessary here. It is deeply unjust that those who, for whatever reasons (chance, tutoring, high quality primary school, etc.) get high KS2 scores are then more entitled to support than those who have identical attainment now, but who once held lower KS2 scores. It would seem to be entrenching pre-existing inequalities in attainment.[11] At child-level, consider, in this model, how you would address the fair question from the student with lower KS2 scores: Why is she [the student with higher KS2 scores] going to intervention and I'm not?

Therefore, at Dixons, we don't want our teachers to worry too much about progress since attainment is the thing we almost always want to know about anyway. As outlined in the previous section, at Dixons, any student towards the bottom of a cohort or class attainment percentile rank will be prioritised for academy-leader-driven or class teacher intervention – in each subject, the goal for all students is to achieve at least a grade 5 at the end of Year 11.

In terms of progress, on starting at a Dixons academy, students are awarded a baseline grade for each subject that they study. These baseline grades are based on a student's KS2 performance relative to their peers. We can use the baseline to calculate value-added (VA) scores after each end of year assessment: *current* approximation of performance minus *baseline* approximation of performance. For example, if a student is awarded a baseline rank '38%' and achieves a percentile rank of '48%' at the end of Year 8, their VA score is '+10%'. If another student is awarded a baseline rank '47%' and achieves a percentile rank of '42%' at the end of Year 8, their VA score is '-5%'.

Of course, individual student grades can be very 'noisy'. This does become increasingly attenuated as data gets aggregated across multiple students and academies. However, underlying differences between individual students' current performance or an individual student's performance over time can either be amplified or hidden depending on their proximity to grade boundaries. And this is before we even consider the various forms of measurement error that might influence any individual student's apparent performance.[6] As outlined in the previous section, we consider individual student grades using a +/- 5% range.

Therefore, we don't share VA scores with students and their families; however, we do use them to recognise those students who have made large shifts from their baseline (≥ +6% VA). Leaders can also ensure that those students who achieve a VA score of ≤ -6% are prioritised for intervention (along with those students with low attainment and those prioritised by class teachers). Using the +/- 5% range, we can be confident that a grade ≥ +/- 6% VA represents a large shift from baseline.

prioritise for leader intervention		recognise for large shift from baseline
VA ≤ -6%	VA -5% to +5%	VA ≥ +6%

Academies may consider reporting progress to students and their families once students sit full past papers that are global in scope and GCSE grades can be awarded meaningfully. This shift in scope usually happens at the end of Year 10; it is curriculum-driven and decided by leaders on cross-cutting teams. At this point Attainment 8 Estimates will be used to derive a baseline grade and students will be measured against the goal to achieve at least a VA of +1 In each subject.

[10] Becky Allen (2018). *What if we cannot measure pupil progress?* https://rebeccaallen.co.uk/2018/05/23/what-if-we-cannot-measure-pupil-progress

[11] Becky Allen (2018). Poor attainment data often comes too late! https://rebeccaallen.co.uk/2018/12/01/poor-attainment-data-often-comes-too-late

For mid-year academy-based assessments, we don't report grades, progress / VA scores or colours; however, by comparing a student's percentile rank on a mid-year assessment to their rank on-entry, it will still be possible to identify those students who have made large shifts and deserve special recognition.

How do class teachers use this information to support students?

For both mid-year and end of year assessments, the focus for class teachers is on ensuring that evidence about learning is used to adjust instruction to better meet students' needs. During academy Data Days, teachers use question-level analysis from summative assessments, alongside students' books, quiz responses and other formative assessment records, to identify and start to plan to address students' critical knowledge-gaps and misconceptions. Bespoke interventions may be targeted at individuals, small groups or the whole class and take place throughout the cycle, usually during curriculum time (in our highly-responsive classrooms and informing the review of SoWs), but sometimes outside of lessons.

Class teachers are also given ranking data for each of their classes to help them to determine which students to prioritise for intervention. Students who are not performing as well as their class peers should be prioritised as well as those who are not on track to achieve at least a grade 5 or VA of +1, whichever represents stronger performance.