Ofqual Board

Paper 6/16

Date:
18 May 2016

Title:
Strategy, Risk and Research Update

Report by:
Michelle Meadows, Executive Director, Strategy, Risk and Research

Responsible Director:
Michelle Meadows, Executive Director, Strategy, Risk and Research

Paper for discussion/information

Open paper (paragraphs 15 & 37 and Annexes B & C closed)

Issue

1. In the past, the Board has received bi-annual reports on the work of the Strategy, Risk and Research (SRR) Directorate. However, to give full visibility of the scope and pace of activities within SRR a paper will now come to each meeting.

Recommendation

2. The Board is asked to note the range of work undertaken within the Directorate and the progress made.

Background

3. We have now published our Corporate Plan for 2016/17 which reflects agreed strategic priorities for the year ahead. The plan and our earlier strategy work have informed business and resource plans for each team in the organisation. They are now being used to set individual's objectives. The Corporate Plan Tracker (contained in the Chief Executive's report) has been updated to match our new commitments.

Analysis

4. We continue to review our regulatory strategy and will publish an updated statement later this year. To inform this, we will present a discussion paper to the Board at the July meeting. This will ask Board members for their view on the role of enforcement versus influence and guidance in our regulation. In particular, the circumstances in which we ought to protect the ability to enforce, accepting the implications for ways of working and for the tone of our approach. A short term Regulatory Strategy Group chaired by the Chief Regulator has been established to help develop the thinking behind the paper.

5. Nonetheless, validity continues to be our primary focus and we are developing our approach to thinking and talking about validity and validation. Our new explication of validity – as the degree to which it is possible to measure what needs to be measured by implementing an assessment procedure – now constitutes a technical point of reference for internal and external discussions. Further, as discussed at the February strategy day, the new approach includes ways of characterising validation, introducing the concepts of macro- and micro-validation to represent lenses of different 'grain size' through which to scrutinise qualification validity. In addition, we have introduced the concept of meta-validation to refer to research and analysis which seeks to unpack what we mean by validity and validation. As a set, these new concepts have been useful for organising work, as well as for discussing the nature of validity and validation with Awarding Organisations (AOs). Indeed, we have begun to disseminate this new approach within AOs, and the ideas have landed favourably with representatives from both general and vocational bodies. Papers have also been submitted for publication in academic journals, to promote Ofqual's approach and to establish its credibility more widely. Further, we have produced a short reference document that sets out the different conceptualisations of validity with examples of each. A draft is attached at Annex A.

Risk

Strategic Risk

6. The strategic risk register, previously included in the CEO update, will now be included in this update. Board members are asked to review and comment on the current risk position as shown in the register attached at Annex B. Progress on established risks is moderate or good, and the direction of travel is neutral across all risks, other than funding pressures on Ofqual, where the direction of travel is positive. With the new Chief Regulator in post, we have closed the related risk.

7. The issue reported to the Board at the last meeting with regards to the Information Management Transformation Project has now been addressed and closed. This is reflected again on the strategic risk register. Revised planning is in place to address the relevant risks –

but the rating remains high. The Chief Operating Officer's paper addresses this in more detail.

8. Board members are asked to review one new risk, related to technical education. The policy development this risk relates to is set out in the Vocational Qualifications update.

Entity Risk

9. We collect and compile data for each entity we regulate to inform judgements of the risk. We have previously committed to issuing entity data sheets to each AO, and intend to do this in June as a one-off trial to see the impact and value. Each risk indicator that contributes to an entity 'profile' has been reviewed to check its justification, accuracy and the likelihood of unintended consequences for AO behaviour. We are engaging with the Federation of Awarding Bodies (FAB) to undertake a pilot issue of the data sheets with a small number of AOs, prior to their wider release. There will be significant communications activity, including FAQs, to ensure that issuing runs as smoothly as possible. We will evaluate the activity later this year to determine the impact, benefits and costs.

Systemic Risk

10. We will bring a systemic risk register, based on the risks presented to the Board in February, to the July meeting. From then on, we intend the Board to review the register bi-annually.

Risk Appetite

11. Our original work on risk appetite was informed by published examples, particularly an approached used by HM Treasury and an example from another regulator. In April we tested our updated approach with a group of internal stakeholders. Despite significantly tailoring the approach to our context, we found that our statements of risk appetite were broadly accurate, but not sufficiently useful to inform strategic thinking and risk management. Particularly, they risked being either too generic to inform good debate; or so comprehensive that they would quickly be out of date, or unnecessarily time consuming to articulate.

12. We have decided, therefore, to take a fresh approach to risk appetite. We intend to develop a series of principles and prompts that would be applied to different situations. These would enable us, on a case by case basis, to analyse different risks in a consistent and principled way – to determine our risk appetite.

13. This year we will refresh our overall risk framework. The risk appetite principles outlined above will be a foundational building block in that work. We will report to the Board later in the year.

Financial Stability

14. We have decided not to routinely collect financial information from every AO because the imbalance of burden to risk. Rather, we will use indictors of market instability to highlight where further data collection may be proportionate, for example, where an AO has had a significant drop in certifications with a marked change in market share. We will judge whether, in these circumstances, it is appropriate to request information about how resources are being managed and whether increased regulatory attention is appropriate given the risk of short-cuts being made.

15. Paragraph redacted – publication would be prejudicial to the effective conduct of public affairs.

Standards

Preparations for summer awarding

16. Since late autumn we have been discussing with exam boards the approach to monitoring in summer 2016 and we are currently finalising the document that sets out the bases of the predictions and the reporting mechanisms. This will be considered at the meeting with Responsible Officers at the end of May and then it will be published on the website as a Regulatory Document. Much of our focus ahead of this summer's awarding has been on new AS. We expect entries in the 13 subjects to be lower than in previous years, as the AS is no longer a requirement for students taking A level, and this might also affect the motivation of those students who are entered. As a result, predictions might be less reliable. We are analysing the provisional entry data and we have also prepared additional guidance material so that exam boards are consistent in the way they make use of the evidence from awarding committees.

17. We expect that schools will see more year-on-year variability in their AS results and we have collected previous years' AS data so that we can publish variability analyses for AS for the first time this summer. We will also publish the provisional entry data (late May) and we will refer to any changes in entry in our open letter to schools and colleges (June).

18. Other priorities over the summer will be the award of CIE IGCSE English, which has been the subject of much criticism from HMC and others in recent months, and A level MFL (see below).

CIE IGCSE English

19. On 11th April, HMC sent us an embargoed copy of a research report written jointly with GSA which they claimed called into question the CIE IGCSE First Language English award in summer 2016. CIE's IGCSE First Language English specification has two entry numbers: 0522 which is available to state schools in England and 0500 which is taken mainly by students in independent schools and overseas centres.

HMC's concerns focused on 0500 which shares the assessments with 0522. We had closely monitored the grading of the specification. We considered the report carefully, and in the light of the evidence we gathered in August and September, we also carried out additional analysis using data that had not been available in the summer. We concluded that the HMC report was fundamentally flawed, in that it was focused on an unrepresentative sub-sample of the entry. Analyses using an unbiased sample did not support the conclusions drawn. Having reviewed the report and the evidence we already had, we did not believe it presented any evidence that called into question the award or our view in August.

20. On Friday 15th April we published a statement detailing our involvement and our conclusion, to coincide with the report's publication. Alongside that we presented a detailed critique of the HMC report, analysis of year-on-year school level variability for IGCSE (which was in line with variability for GCSE English), and a Rasch analysis that showed that the difficulty of CIE's IGCSE English was in line with all exam boards' GCSEs in English. These can be found at https://www.gov.uk/government/news/cie-igcse-first-language-english-summer-2015

21. *This paragraph has been redacted as its publication would be prejudicial to the effective conduct of public affairs.*

A level French, German and Spanish

22. Stakeholders have expressed concerns that some of the issues identified in our September 2014 report were not addressed in the 2015 assessments. We have carried out further work, including reviewing the analyses carried out by the exam boards, and reviewing a number of summer 2016 A2 papers. We are also surveying schools and colleges entering students this summer (approximately 2400 centres) to gather information on a number of 'native speaker' features. This will enable us to estimate the proportion of students that might be native speakers, or have some experience gained outside school, and we will test whether the presence of these candidates has an impact on the reliability of awarding data.

Additional analysis of summer 2015 EAR data

23. In the autumn we committed to collecting more detailed data on EAR. Our analysis has taken longer than expected, largely due to the poor quality of the data submitted by exam boards. We expect to have published the outcome of these additional analyses by the time of the Board meeting.

24. Headline findings are as follows: Independent schools in general (but not always) have a higher rate of EAR as a percentage of their overall entry and more so at A level than GCSE. There is no obvious pattern of

grade changes by subject/level but unsurprisingly there are more grade changes in the more subjective subjects. At GCSE the most commonly challenged grade is D (37% of EAR overall, nearly half of English and maths EAR). At A level the most commonly challenged grade is B. Average mark change varied by subject and is unsurprisingly higher in the more subjective subjects, but it is never more than 3 raw marks.

25. We also looked at EAR resulting in changes of two or more grades. Only 639 qualification grades out of the 8 million awarded qualifications were changed by two or more grades. Perhaps surprisingly, changes of this magnitude were most common at AS, then GCSE then A level. A handful were downward changes. Many of the A level changes were grade B to A* where the candidate gained one additional A2 mark to get the 90% of A2 and they already had enough UMS for an A. The most common reason for the grade change given by exam boards was misapplication of the mark scheme although the boards were not good at providing detailed reasons for the changes.

26. Progress 8 as a measure of secondary school accountability is implemented this year. Because it is based on a measure of student progress from Key Stage 2 attainment in 8 GCSEs, it is likely that the key focus will shift from grade C to all grades. This may be reflected in EAR where historically a substantial proportion have come from C/D borderline candidates. We are just beginning the analysis of data from 300 schools who opted into progress 8 early in 2015 to gather some insight into the likely effect on EAR. We will share the findings with boards to help their planning and we will also communicate more widely so as to prepare stakeholders for changes in EAR patterns.

Research

Annual Perceptions Survey

27. The Board has expressed concern over the publication date of the Perceptions Survey because any negative media attention it receives can be unsettling for those involved in exam series. With this in mind we have explored whether it would be possible to delay publication. The survey is classed as official statistics. This means that once the report is ready it must be published as soon as possible. Hence, it will not be possible to avoid publishing within the exam series this year. We are however carefully considering exactly when to publish and we are bringing forward next year's survey so that it will be published in the early spring. We are also reviewing the content of the survey to ensure that it is as informative as possible. Changes were also made this year, for example to the wording of questions to make clear that exam boards are responsible for the quality of marking.

A level Science Practical work

28. Our evaluation of the impact of the removal of practical work from the A level Science grade comprises a number of strands. One involves testing university students' ability on specifically designed practical skills tasks, comparing students who have sat current specifications

with those taking reformed specifications. We are about to pilot the practical skills measures in each of the three sciences to ensure that they are manageable and valid.

29. We are also interviewing science teachers to understand the impact of the changes on their teaching and how they are managing the new arrangements. Early indications are that while there are some uncertainties about details of the new model, such as the extent to which tasks can be varied, or the amount of evidence needed for the endorsement, teachers are generally positive about the new arrangements and crucially, they are not having a detrimental effect on practical skills activities.

30. Other bodies are also conducting research to evaluate the policy change. We are working with a group of stakeholders keen to monitor and support the implementation of reforms. They include the Royal Society, charitable organisations (including Gatsby, Nuffield and Wellcome), exam boards and others. It was agreed, in 2015, that Ofqual would draft a proposal for an independent project to keep track of the various sources of evidence that would be produced over the coming years, to understand where problems might be arising and to support conclusions concerning the success of the reforms.

31. The draft proposal, emphasises the importance of independence, rigour, and adopting a broad focus, to ensure the credibility and usefulness of outcomes. It was envisaged that the evaluation would not report until some years after the first examinations, but that there would be interim products along the way, including a repository of evidence and a report on the background to the policy decision. The proposal was received well by many of the stakeholders, but not all. The charitable organisations were not convinced that the proposal was worthwhile funding. We are currently in the process of deciding how to respond to these concerns, before developing another version of the proposal.

Standard setting and maintaining in national examinations

32. This is a new joint project between Ofqual, Oxford University and AQA. Its focus is national, curriculum-related exam systems from a wide range of jurisdictions around the world. Experts from about 15 jurisdictions are being recruited. Their first task will be to describe the processes used to set or to maintain standards in their jurisdiction and to explore the conceptualisations of standards that lie behind these processes.

33. We then plan to hold an international symposium (March 2017) at which experts will discuss the descriptions – comparing and contrasting approaches. The final outcome of this project will be a book that will be the first of its kind, as it will document the diversity of approaches taken to standards internationally. It is anticipated

that the project will challenge current theoretical positions on standards, give us a broader context in which to view our system and may expose good practices from which we can learn.

Ofqual's international standards objective

34. Of course, the latter work does not address the international element of our standards objective. The qualifications standards objective is to secure a consistent level of attainment (but not over time) between regulated qualifications and comparable qualifications awarded outside the UK.

35. We wish to ensure that we meet the international element of our standards objective in a proportionate manner and have begun to set a strategy for doing so. We published a substantial research project comparing A levels with international equivalent qualifications in 2012. The findings provided a context for the following A level reforms. For GCSE, we conducted international comparisons of the demand of question papers in maths in 2015 but have not yet fully mined those data. Our thinking is, therefore, that work on GCSE is of higher priority than A level and that the first awards of the new qualifications is the point at which research work would be most helpful. We are also exploring the use of international survey data as a crude linking mechanism but richer (yet narrower) pieces of research on specific qualifications would be informative.

Quality of Marking Metrics

36. We have now calculated a range of metrics and graphical displays of quality of marking at unit, specification and subject level for GCSE and GCEs in English language, English literature, French, Spanish, geography, history, business studies, physical education, psychology and physics, for the past 3 years. In total, the data represents 23.5 million items.

37. Paragraph redacted – publication would be prejudicial to the effective conduct of public affairs.

38. The metrics have been shared with boards and to date feedback has been positive although they have expressed concern about the potential negative consequences of regulatory use or publication. Indeed, the complexities and sensitivities around the data (for example, the risk of information being taken out of context) and, critically, the importance of the metrics not compromising exam boards' live monitoring behaviours, make publication risky. Indeed, we are currently considering how to use the metrics as a positive tool in regulation (for example, sharing board specific information or identifying subject/units in need of additional monitoring). Some of the findings and issues with using the metrics are expanded upon in Annex C.

GCSE Science and A level Mathematics accreditation

39. To support the accreditation of GCSE Science two rounds of a comparative judgement exercise were conducted in which science teachers compared the difficulty of pairs of items from the sample assessment materials and from the current exams. Whilst this work has helped support accreditation decisions, it has also raised some interesting insights into judgemental biases in teachers regarding item difficulty. For example, multiple choice items tended to be judged as less difficult than students actually found them and items requiring calculations were judged as more difficult than they actually were. We were able to correct for such biases in the data. Now that the science specifications are accredited we will publish a research report on this work.

40. It will be possible to test the generalisability of these findings as we are now supporting the accreditation of A level mathematics in a similar way. However, A level mathematics items are more complex than items included in previous studies. For high tariff items the method for allocating intermediate marks will play an important role on overall item facility. We are therefore conducting pilot work to test the impact on the reliability of judgements of including the mark scheme, different types of judges (teachers and PhD students) and so on.

Measuring the sawtooth effect in GCSEs and A levels

41. It is believed that when qualifications are reformed, student performance drops in the first year of testing and then improves over subsequent years as teachers and students become increasingly familiar with the assessment arrangements and as materials (e.g. past papers) become more available. This is commonly referred to as the sawtooth effect. Last year we found evidence to support the notion of the sawtooth effect in the pattern of grade boundaries awarded. This is important because it may help us interpret NRT outcomes. This is being followed up with a judgemental strand in four subjects: mathematics, history, geography and religious studies using an adaptation of the comparative judgement technique comparing quality of work in candidate scripts. This work is currently underway.

Research Advisory Group

42. Our programme of research is now at a level where it would benefit from external scrutiny. We are in the process of setting up a Research Advisory Group to be chaired by a member of the Board (Mike Cresswell). The group will consist of between four and six external experts. In the main, the experts will have significant research experience in the fields of education, assessment and qualifications. The group will meet bi-annually and will advise Ofqual's Executive Director for Strategy, Risk and Research on the quality of research papers as well as methodological, design and analytical considerations of planned research. The group will not overlap with the terms of reference of the Standards Advisory Group.

Paper to be published	YES except paragraphs 15 & 37 and Annexes B and C
Publication date (if relevant)	After the meeting
If it is proposed not to publish the paper or to not publish in full please outline the reasons why with reference to the exemptions available under the Freedom of Information Act (FOIA), please include references to specific paragraphs	Publication of the sections identified would be prejudicial to the effective conduct of public affairs

ANNEXES LIST:-

ANNEX A Micro-validation, Macro-validation, and Meta-Validation

ANNEX B Strategic Risk Register (closed)

ANNEX C Quality of Marking Metrics – Some Issues With Use (closed)

ANNEX A

Micro-validation, Macro-validation, and Meta-Validation

Ofqual has recently adopted a distinction between micro- and macro-validation, to indicate different 'lenses' through which to scrutinise an assessment procedure in order to judge whether it demonstrates sufficient validity.

Macro-validation, at one end of a continuum, describes a 'lens' which is wide but lacks detail. It tends to focus on 'high level' evidence of validity – e.g. patterns in results, or a public confidence survey – and asks whether that evidence is consistent with the claim that it is possible to measure what needs to be measured by implementing the assessment procedure in question.

Micro-validation, at the other end of the continuum, describes a 'lens' which is narrow and therefore highlights detail. It tends to focus on 'low level' evidence of validity – specifically, the nature of the features and processes which comprise an assessment procedure and how they operate in practice – and asks, in relation to each feature or process, whether it appears to have been effectively designed.

This distinction suggests that, for any particular assessment procedure, you can undertake validation either holistically (macro-) or atomistically (micro-). Ideally, validation ought to combine both perspectives. The distinction is intended to foreground the importance of atomistic validation techniques, which have traditionally not been foregrounded in the literature.[1]

Ofqual has also adopted the term **meta-validation**, to describe research and analysis into validation itself, rather than into the validity of any particular assessment procedure. This would include work on the nature of validation, e.g. the identification of important threats to validity throughout the qualification lifecycle. It would also include work on validation standards, e.g. the identification of criteria for distinguishing between sufficient and insufficient validity in a particular regulatory context.

[1] In other words, the precise allocation of a particular technique to one 'lens' rather than another is not critical.

Additional examples

Micro-validation

The correlation of results from a newly developed assessment with results from an established ‘yardstick’ assessment (whose results you already trust as a measure of the proficiency in question) provides a good example of a macro-validation technique. By using a ‘wide angle’ lens it allows you to draw an inference about the validity of the assessment procedure overall. So a high correlation (from this one technique alone) provides evidence that it might well be possible to measure what needs to be measured by implementing the assessment procedure in question – in this case, the newly developed assessment. A low correlation provides evidence to the contrary; but provides no ‘diagnostic’ information as to what might be wrong with the assessment procedure.

Micro-validation

The ‘cognitive laboratory’ technique, for investigating how candidates interpret what each of the questions posed by an assessor is asking them to do, provides a good example of a micro-validation technique. By using a ‘telephoto’ lens it allows you to draw an inference about the effective design of a particular feature or process within the assessment procedure – in this case the wording of particular questions – but typically does not allow you to draw an inference about the validity of the assessment procedure overall. So evidence that certain candidates, who might otherwise be expected to answer a particular question correctly, are misdirected by a picture which accompanies it provides evidence that the particular question is not making its intended contribution to the validity of assessment procedure overall. Clearly, though, with such a detailed focus, there is much ‘diagnostic’ information available with which to understand and rectify the situation.

Meta-validation

Examples of meta-validation might include:

- analytical work on an approach to thinking and talking about validity and validation
- a literature review of good practice in linking examination standards across similar qualifications
- empirical research into levels of reliability that can realistically be expected of certain kinds of vocational qualification.