Setting and maintaining standards in national examinations around the world

Professor Jo-Anne Baird, Director of the Oxford University Department of Education, England jo-anne.baird@education.ox.ac.uk (presenter)

Dr Lena Gray, Head of Research in the Centre for Educational Research and Practice at the Assessment and Qualifications Alliance (AQA), England lgray@aqa.org.uk

Dr Tina Isaacs, Senior Lecturer at the UCL Institute of Education in London, England Tina.Isaacs@ucl.ac.uk

Dennis Opposs, Standards Chair, the Office of Qualifications and Examinations Regulation (Ofqual), the exams regulator for England dennis.opposs@ofqual.gov.uk (presenter)

Abstract

Assessing the achievement of curriculum standards is powerfully enacted through processes of standard setting and maintaining within curriculum-related examinations. Many countries use curriculum-related examinations to select learners for higher education, work and other study options. Some countries also use these examinations as tools to measure school system performance; as such, these examinations shape the landscape of senior school education, defining education system quality for learners and for society.

Given the high stakes nature of these examinations, it is surprising that the ways examination standards are conceptualised and operationalised differently across nations has not been given sufficient attention. This is an interesting area because globalisation has begun to impinge on examination systems, but public examination standards are still largely a bastion of the local. The meaning of ‘standards’ differs between countries and the stated value positions and processes relating to examination standards differ markedly. How policy and politics affect standards in different countries has not been articulated well. Further, there is a tenuous relationship between standard-setting theory and the manner in which standards are operationalised by nations.

This paper reports on phase one of a project that aims to examine critically policy positions and processes for assessment standards in a range of countries, drawing on analyses from in-country experts and researchers. The project aims to illuminate similarities and differences in conceptual bases and operational approaches to standards. It is anticipated that it will challenge current theory on standards, and lead to changes in how national organisations approach standard-setting and maintaining. The paper will be of interest to researchers, policy-makers and practitioners interested in assessment standards. Discussion with colleagues at the conference on the extent to which our initial findings contrast or coincide with views of assessment standards in their context will help in moving forward.
Keywords: curriculum standards; examination standards; assessment standards
**Introduction**

The focus of the project is national, school-leaving or university entrance, curriculum-related examination systems from a wide range of jurisdictions around the world. The project aims to describe the processes used to set or to maintain (or link over time) standards in these examinations and to explore the concepts relating to standards behind them. (This paper uses the phrase “standard-setting systems” to cover both systems for setting standards and systems for maintaining standards.)

This project is needed because there is a lack of accessible documentation at a level that could be interpreted in terms of the information used, emphasis given to various sources, acceptable ways of defining standards and issues that would lead to controversy. Such a project would add to our knowledge and would help us all to view our own examination systems in a broader context. It could be a very important project for those responsible for these types of examinations in a range of jurisdictions and contexts.

The first output from the project will be an international symposium in Oxford in March 2017 at which experts from around the world will come together to discuss papers that they have written about the systems they know. Presently we are aiming to include papers from approximately 15 jurisdictions at this symposium.

The final outcome of the project will be a book that will be the first of its kind, since it will document the diversity of approaches taken to standards in different jurisdictions internationally. These jurisdictions will be selected from those discussed at the symposium. A selection of the papers written for the symposium will form the basis for several of the chapters in the book.

No individual presently has access to the breadth of information that this project will produce, so it will truly be ground-breaking. It is anticipated that the project will challenge current theoretical positions on standards, as well as lead to changes in how organisations approach standard-setting systems.

Secondary gains we can see from the project include establishing an international network in this field, from which other projects and useful connections could arise. The project findings may, for example, provide a useful resource for those working on international comparability projects, which are becoming increasingly common and high profile.

We look forward to discussing these issues with colleagues from different jurisdictions and to exploring the benefits of this project for the international assessment community.
Key issues for the research project

The research aims for the whole project are to investigate, document, analyse and evaluate four key aspects of national standards-setting systems:

- How standards are defined in national examination systems
- How those definitions are enacted in terms of processes and evidence used
- Issues for the system and responses to those issues
- The commonalities and diversity of definitions of, processes for and challenges to standards.

How standards are defined in national examination systems

Very little has been written that documents and conceptualises the meaning of examination standards in high stakes national examinations. In England, although the meaning of examination standards has been much debated in the literature (Baird, 2007; Baird, Cresswell & Newton, 2000; Cresswell, 1996; Christie & Forrest, 1981; Coe, 1999, 2007, 2010; Newton, 1997a, 1997b, 2003, 2005, 2010), it is often noted that stakeholders discuss examination standards using contradictory definitions, often without realising they are doing so.

When curriculum-related examinations are used for high stakes purposes such as university entrance, a number of comparability challenges arise, for example between subjects, between examination boards or over time. Given the importance of these examinations for people’s life chances, it is vital that we try to understand these comparability challenges. There is a yawning gap in the academic and policy literature that makes it difficult to understand examination systems internationally and to explain how they operate, how they are interpreted and why certain issues are a matter of public controversy in one country but not in another. Transparency of the meaning of examination standards matters to educational professionals as well as to candidates, parents, politicians and the media. Each may have their own definition of what it means for standards to be ‘comparable’.

This project seeks to document and compare how standards are defined in a range of jurisdictions around the world. In doing so, it looks to understand the similarities and differences in the conceptual bases for standards and to analyse how those conceptual bases relate to the standard-setting systems used, and to the meanings and values ascribed to standards.

How those definitions are enacted in terms of processes and evidence used

A second area of focus for the project will be the processes used to set and maintain standards, including the personnel involved and the evidence used in the process. Most national examination systems use both statistics and examiner judgment in their standard-setting processes (Cizek & Bunch, 2007, p. 10), but lack of transparency regarding how
various sources of information are utilised in decision-making is common. Baird et al. (2000) argued that examination boards had to gauge the values of stakeholders in setting standards and that the prioritisation of various sources of evidence was essentially subjective. In part, this was, they argued, due to the fact that no single source of evidence adequately captured the meaning of examination standards.

The relationship between the standards definition, the evidence and processes used, and the challenges that the examination system must be able to defend against, is a complex and fascinating one. In some contexts, employing different sources of evidence, a range of personnel, and some diversity of processes, may allow examination systems to be able to defend themselves against challenges from a range of different perspectives. In other contexts, there may be strong challenge from a fairly uniform perspective, and the system response may be to include a number of checks and counter-checks to try to ensure that the defined problem does not occur.

It is these issues that the project seeks to explore by documenting standard-setting processes, and it is hoped that in doing so the project will provide a resource which policy-makers, regulators and examination boards can use in order to compare the methods used in their own system with those used elsewhere in the world.

**Issues for the system and responses to those issues**

This project is not about documenting standard-setting systems in order to judge or rank them. Examinations and their associated standard-setting systems are culturally situated within an education system, which is in turn shaped by prevailing social values. Whilst we do not seek to make value judgements, aspects of a public examination system that are a matter of public debate can be very telling in helping to illuminate the challenges to and limitations of the standard-setting system, and consideration of those public issues and challenges is therefore a key part of the project. Our aim in documenting these issues and challenges is to use them to help analyse any differences between intended and enacted definitions and processes, and the underlying reasons for these differences.

Documenting these challenges and responses should, we hope, allow us to discuss common features in ways that will be helpful in future to those designing or redesigning standard-setting systems for a particular context.

**The commonalities and diversity of definitions of, processes for and challenges to standards**

Documenting, comparing and analysing how standards are defined in national examination systems, how those definitions are enacted in terms of processes and evidence used, and the
issues for the system and responses to those issues should allow us to provide some truly
ground-breaking research that evaluates the commonalities and diversity between different
systems. Comparative work will focus on aspects of setting and maintaining standards such as:

• philosophy and processes
• issues and problems, including backwash effects
• who “owns” the standard
• identification of different practices.

In doing so, comparative work will take into account:

• the theoretical basis for standards
• methods used to set and maintain standards, including the triangulation of
  information sources
• decision-making processes
• social issues such as power, authority and trust.

Methodology

One of the key issues to be addressed in this project is what research techniques we can use to
ensure that we have the most accurate and fullest picture of any standard-setting system. The
project uses a case study methodology, and overall could be described as a multiple-case
embedded model (Yin, 2014, p50), that is, it involves multiple cases, each with its own
contextual conditions, but with multiple units of analysis within each case. We selected this
methodology because, in Yin's words,

Case study research would be the preferred method, compared to others, in situations
when (1) the main research questions are "how" or "why" questions; (2) a researcher
has little or no control over behavioural events; and (3) the focus of study is a
contemporary (as opposed to entirely historical) phenomenon. (Yin, 2014, p2)

Our use of a multiple case approach is intended to "shed empirical light about some
theoretical concepts or principles" (Yin, 2014, p40) by comparing cases that mirror and
confirm existing documented definitions of standards and those that contrast, and so provide
a challenge to those documented definitions and allow us to move thinking forward. Our
chosen cases, then, have been selected to illuminate distinctive approaches to standard-
setting, geographical spread, cultural distinctiveness, developed/developing jurisdictions,
different assessment formats and use of differential cut-scores from the same examination.

Having selected a case study method and a multiple case design for our study, our next
methodological challenge was to try to establish what our sources of evidence would be.
Analysis of documents and archival records will form part of each case study, embedding
different units of analysis within each case study. A pilot study on standard-setting in
Scotland and England showed us that documentary and archive evidence is not enough on its
own, as publicly documented positions on standards can be either brief, or indeed may be contradictory, either within the same source or across different sources. We needed, therefore, to use multiple sources of evidence (Baird and Gray, 2016). We considered interviews and direct observation as possible additional sources, but given the international nature of the project, these two sources of evidence are likely to require people and budgetary resources that we don’t have. Additionally, these methods were likely to suffer from lack of depth of understanding of the educational and assessment environment on the part of the researchers. We therefore decided to use participant observation as our main source of evidence, using participants who already work within the system under investigation. This method provides for immediacy of data, and gives us access to evidence which may not otherwise be available, as Yin points out:

Participant observation provides certain unusual opportunities for collecting case study data but it also involves major challenges. The most distinctive opportunity is related to your ability to gain access to events or groups that are otherwise inaccessible to a study. (Yin, 2014, p116)

Using participant observers who are already part of the jurisdiction under investigation means that we need contributors who know the system well, and who are in a position to discuss publicly and document issues that could be controversial in their own context. For the most part, we are selecting senior personnel in examination boards to write the country-specific papers and chapters. This does give us a risk of potential biases, though, and in particular creates a threat that findings are limited in a number of ways. One of the most important is that the papers may cover the policy intentions, rather than the actual system. Even in setting out the “official” position, there may be variations in how full a description can be provided in each country-specific chapter. Some systems will have complete policy statements, and perhaps public documents and archives on standard-setting processes, but in other systems, this information may not be in the public domain, and an examination board employee may not be in a position to release it into the public domain.

Moreover, as Baird and Gray suggest, “Examination boards have a tricky, political task in managing public and stakeholder perceptions of examination standards” (2016, p2) and therefore another limitation may be that examination board personnel may not be in a position to discuss public critiques of their system in a full and open way. Senior examination board personnel are powerful people in the sense that they are "those with great responsibility and whose decisions have significant effects on large numbers of people" and therefore any research involving them is sensitive because it is likely to encroach upon "issues about which there is high-profile debate and contestation" (Cohen, Manion and Morrison, 2007, p127). No matter our research methodology, confidentiality cannot be assured to participants: the participants (or in this case, their organisations) are high-profile and identifiable. Indeed, it is central to our research design both that the organisations should be identifiable and that the research should foreground and analyse issues that are subject to public debate.
We know from comparative work in education that the policy and educational reform landscape is constantly moving in many jurisdictions around the world (OECD, 2015). Such policy shifts may create space that allows discussion of the strengths and weaknesses of different systems, processes and concepts of standards: on the other hand, policy reform can lead to "resistance from policy-makers to listening to the concerns raised by education and assessment professionals" (Baird and Coxell, 2009, p114), stifling debate and making it difficult for professionals like examination board employees to articulate their knowledge in public.

Despite these tensions, there is little doubt that senior examination board personnel are likely to have the knowledge to provide a full and accurate description of their system, its underlying principles, and how it works in practice. In order to mitigate the risks of biases in research data gained from participant observation, we need to take extra care to test the quality of the findings and to check the validity and reliability of the research. In order to ensure that our case studies are significant, the project needs to include alternative perspectives:

An investigator must seek those alternatives that most seriously challenge the assumptions of the case study. These perspectives may be found in alternative cultural views, different theories, variations among the stakeholders or decision-makers who are part of the case study, or some similar contrasts. (Yin, 2014, p204)

A key way that we will do this is by ensuring that the project addresses rival explanations of the phenomena described, by asking additional in-country experts to provide commentary and analysis on standard setting and maintaining policies and processes within the jurisdiction. We will also provide alternative perspectives through overarching chapters that draw out key themes from across the case studies, as described above.

At the time of writing, in-country experts are being recruited and symposium papers starting to be drafted.

**Conceptualising examination standards**

While the acts of setting and maintaining standards seem to be distinct – one is ostensibly about what we do when faced with a ‘new’ specification, syllabus or examination, and the other is about keeping what we’ve done in line with what we’ve done before – in practice setting new examination standards does not happen in a vacuum, but is usually done with reference to standards we’ve previously set on similar specifications/examinations. Indeed, both popular debate and the literature on the meaning of examination standards reflect attempts to maintain – and compare – standards across examinations and over time.
In the UK, which arguably privileges standards over time above other considerations, much of the literature about examination standards during the past few decades has revolved around defining standards in a way that encompasses not only the act of situating standards but keeping them stable. Many of the articles have put forward particular definitions (e.g. Baird, 2007; Baird et al., 2000; Coe, 2010; Cresswell, 1996; Wiliam 1996). In 2010 Newton tried to come to grips with these definitions and focused on examinations takers’ attainment as the common thread. He put forward three ways of categorising standards that revolved around what the students who have achieved ‘comparable’ standards have in common. He labelled as phenomenal standards those definitions that rely on attributes of students’ performances, where question demand or the answers that students give on examinations could be used as evidence that comparable standards have been maintained. A causal definition is one in which we look to have students with similar background characteristics (for example prior attainment) gain similar results. The third, predictive, category would expect that students with similar outcomes could progress to similar achievements. These categories say little about methodology – for example, expert judgment on attainment could be used in all three, as could statistical evidence. We have taken on board Newton’s categorisation because of its focus on attainment factors and its ability to incorporate previous definitional approaches, but we are aware that this is contentious. Our pilot study on standard-setting in Scotland and England showed that there was a superordinate category of ‘attainment referencing’, which was not bounded by Newton’s categories. Through exploration of definitions of standards in other jurisdictions, we will be able to categorise approaches and are open to the potential requirement for extension of our working taxonomy.
References


