Norwegian Knowledge Centre for Education
(case number 13/4697)

State of the Field Review

Assessment and Learning

Jo-Anne Baird,[a] Therese N. Hopfenbeck,[a]
Paul Newton,[b] Gordon Stobart[b] and Anna T. Steen-Utheim[c]

[a]Oxford University Centre for Educational Assessment
[b]Institute of Education, University of London
[c]BI Norwegian Business School and University of Oslo

Oxford University Centre for Educational Assessment Report OUCEA/14/2

KNOWLEDGE CENTRE FOR EDUCATION

UNIVERSITY OF OXFORD

Contents

Chapter 1 Executive summary

The state of the field review of the research literature on assessment and learning covers major developments internationally this century. The intersection between assessment and learning is of utmost importance for the promotion or hindrance of quality in education. This topic is important around the world, but Norwegians have specifically been calling for more work in this area so that questions can be answered about how much children are learning in schools. As in other countries, there has been concern that there is a lack of an assessment culture. For a long time, grading children's schoolwork was seen as detrimental to their learning, motivation and feelings about themselves. Cultural and historical reasons for this differ in each setting, but in Norway this is related to societal values of equity, inclusion and democracy, which were antithetical to national testing. In this environment, the introduction of international tests provided hitherto unavailable sources of data. Our cross-cutting theme in this review relates to how valuable learning is shaped and defined by assessment theory and practice. How societies, including Norwegian society, will be influenced by the uptake of particular approaches to defining learning through assessment technologies[1] is an important issue. Assessment is not a neutral activity and through this review we seek to depict how the literature has explored the relationships between assessment and learning.

1.1 The rise of assessment

Assessment has assumed much greater importance this century due to its use in new, neoliberal public management systems and structures. First, there is much discussion of the knowledge economy in which economic competition between countries is no longer only on the basis of raw resources and manufacturing, but is also in terms of who can produce the most and best knowledge workers who will take charge of large and international companies. Therefore measuring an individual's, institution's or society's currency in this economy is important. Second, setting targets and measuring productivity is a vestige of new public management that is fed by assessment data. Actors within these assessment data systems, with targets to meet, find themselves with professional and ethical dilemmas because the assessment results and the stakes attached to them can come to be seen as more important than the entity that they signify: the learning. Hanson (1993, 2000) predicted that the signifier (assessments) would be seen as more important than the signified (learning). Even if this prediction has not come true in Norwegian society today, this is the context of the research literature in this area and explains the importance of assessment in education internationally.

[1] Such as classroom assessment practices or large-scale testing based upon psychometrics.

1.2 Narrative review method

The review was designed to depict the state of the art in research on assessment and learning in the 21st century. In addition to a broad outline of the relationship between learning theories and the development of assessment and psychometrics, we looked at the trajectory of validity theory, as it is the central concept in assessment. Two large and growing areas of research activity in this field this century were identified, which are significant also for Norway: international testing and assessment for learning. We explored connections between assessment and learning throughout the report, identifying alternative theoretical perspectives.

To support this narrative review, we conducted systematic searches of the literature on international tests and assessment for learning. Over 850 articles on international tests and over 900 articles on assessment for learning were classified systematically so that we could identify those most relevant to the themes of our review. We used this systematic approach to identify any key articles that might have been overlooked had we relied purely on our knowledge of the field. Additionally, we documented the doctoral work in the area of assessment conducted over the past 20 years in Norway.

The choice of international tests and assessment for learning as areas of focus was interesting because they are both powerful forces in the literature, but are distinctive. International tests are focused at the system level, involve psychometric technology and have no direct impact upon teachers and learners. Assessment for learning is at the classroom level, involving a wide range of (usually) teacher-selected or teacher-made assessments, and is closely connected with direct student learning through feedback and other mechanisms.

This is a narrative critical evaluation review based upon the knowledge and judgement of the authors. To maintain the coherence and integrity of the evaluation, the authors regularly discussed the development of the ideas and the writing. A symposium was held on a preliminary report and the authors took into account the discussions and issues raised. Further, an advisory group was consulted on a draft text and its comments considered in the finalisation of the report. Our advisory group members were Professors David Andrich, Mary James and Dylan Wiliam. Clearly, the views expressed in this report are those of the authors and may not reflect those of the advisory group.

1.3 Learning theory and assessment

Jo-Anne Baird

Assessment has been distinguished from the field of psychometrics, which it subsumes, by authors such as Desmond Nuttall (1987). One of the features of this field is the differing extents to which the assessment and psychometrics paradigms have impacted upon particular national psyches and practices. Arguments for an assessment paradigm, as separate from psychometrics, were closely connected with views about learning and the interaction between assessment and learning. The desire to produce authentic assessments based upon the context of real-life performances is closely connected with this literature. This is also relevant to the extrapolation problem outlined in the chapter on validity (Chapter 8). Caroline Gipps (1994) wrote that in the assessment worldview, learning was developed, criterion-referencing was used, validity was the central concern and the context of performances was designed to allow individuals to show what they know and can do. This was contrasted with psychometric approaches which saw learning as a fixed property of the individual's ability, were norm-referenced and treated reliability as the central concern (Gipps, 1994).

Some authors have depicted the trajectory of theories of learning and related them to changes in assessment design (eg Shepard, 2000; Elwood, 2006; James, 2006). This is very difficult to do because learning theories and different assessment practices have coexisted chronologically. In this report, we categorise learning theories as behaviourist, cognitive constructivist or sociocultural. Establishing any assessment practice as behaviourist, for example, is hard to substantiate because, at least with the benefit of hindsight, few assessments really appear to be unconcerned with mental activity. Behaviourist theories of learning were explicitly uninterested in mental processes. As such, unchanged since Pellegrino et al.'s (2001) review, the state of the art in assessment practice appears to be related to cognitive theory. We see this too in Stobart's (2014) book on *The Expert Learner*. It is possible to construe this as a time lag between developments in learning theory and its impact upon assessment practice, as application of theory is challenging and it can take decades for the implications of theoretical advances to be worked through to practice. However, the challenges of applying sociocultural theory seem to be more fundamental than this. Sociocultural theory, we argue, does not sit well with the current state of the field of assessment practice, in large part because standardised, same-for-all assessments have been equated with fairness in the minds of many. There is a range of interesting work that is trying to resolve the tensions in this area. However, the state of play in the field of assessment is based upon cognitive constructivist approaches.

1.4 Assessment for learning

Gordon Stobart and Therese N. Hopfenbeck

The origins of formative assessment and assessment for learning (AfL) are traced in the literature, as well as their connections to learning theories. This is a complex task since definitions continue to broaden and emphasise change, which leads some authors to shift their theoretical orientation, for example moving to a more sociocultural stance. Our systematic search of the literature on AfL since the 1998a Black and Wiliam article included investigation of whether there was empirical evidence of impact on learning in the form of effect sizes. Whilst claims for large effect sizes are regularly made in the literature, the evidence for these has increasingly been critiqued, for example by Bennett (2011) and Kingston and Nash (2011).

Our systematic review showed that the vast majority of studies on AfL are small-scale action research designs and are published in a wide range of journals. A concern for the review is that current definitions of formative assessment/AfL cover a wide range of teaching and learning practices while research designs often lack an action theory (what is causing change), often accompanied by a lack of systematic data collection (for example baseline data before a research initiative).

Our overall conclusion is that the effects of formative assessment upon learning have been over-sold by some authors. This is unfortunate because the limited empirical research suggests a modest, but educationally significant, impact on teaching and learning.

1.5 Psychometrics and learning

Jo-Anne Baird

Psychometrics may have developed from a desire to study mental events, but it is essentially a field of statistical techniques and is concerned with the outcomes rather than processes of learning. Psychometricians have expressed disappointment with the paucity of uptake of their sophisticated techniques in the literature and have been concerned that a lack of statistical skill could explain this. Given the largely qualitative nature of much research in education, there is a point here, but this is not the only reason. The construct being assessed defines the domain and should also indicate what constitutes progress in that domain (learning). It is remarkable, therefore, that we do not see a large contribution of the field of psychometrics to learning theory. Equally, Sijtsma (2006) noted that psychometrics barely uses learning theory to underpin test design. Whose role it is to provide theory and the kinds of theory sought are debated in the literature. Some see theory as mathematical models underpinning the psychometrics, whereas others are looking for educational theories. Explanations

offered by these different types of theory are bound to differ and thus fall short for the opposing camp.

Nonetheless, it could be said that our knowledge of learning really should have gained more from the systematic collection of data from assessment and testing – whether or not we come from an assessment or psychometrics paradigm. The trouble is that there are underlying tensions about how learning is being defined by the assessment technology used which persist because education expresses values and assessment credits the display of some of those values. Assessments utilising a psychometric approach come with technological baggage relating to how tests are best constructed, analysed and operated. If your values do not align with those practices and you are instead more interested in assessment of learning during classroom interactions, for example, then psychometric approaches are likely to jar with your professional beliefs and practices.

1.6 International tests

Therese N. Hopfenbeck and Jo-Anne Baird

Conceptions of learning in international tests are rooted in the cognitive model of learning. Underlying definitions of learning differ between international tests, although the differences are not so much related to learning theories as conceptualisations of knowledge. Trends in International Mathematics and Science Study (TIMSS) is a curriculum-oriented test, which attempts to assess students' knowledge of items relevant to the school curriculum they have studied. The tests of the Organisation for Economic Co-operation and Development (OECD) (eg Programme for International Student Assessment (PISA), on the other hand, are literacy-oriented approaches, which claim to be curriculum independent. One difference in practical terms between the two approaches is that the PISA items have a higher reading load, which can obviously disadvantage students with low reading abilities. Nardi (2008) showed how interpretations of the outcomes of the tests were dependent upon their relationships with the national curriculum – whether they were commensurate with it, went beyond it or under-represented it. If the tests went beyond the national curriculum, students were tested on material that they had never been taught. The results might be interesting, but any policy questions would relate to the appropriateness of the design of the national curriculum, which would not be informed by the international test *results*. If the international tests under-represented the national curriculum, students would have a wider knowledge of the subject than the tests represented. In this case, a country's ranking would not provide particularly useful information, since that country had presumably already decided that a wider (eg science) curriculum should be taught in schools than was the view of the creators of PISA.

There is not a great deal of evidence of the direct impact of international tests upon learning as yet. However, the literature does show that some countries have aligned

their curricula and tests to the design of PISA. Therefore it is possible that we will see a greater impact of international tests upon learning directly over the coming years. However, looking at international tests only through a student level lens is to miss some of the most important effects that they have, as they are designed to look at educational systems and this has implications for their use as well as their assessment technologies. A section of the research literature has looked at the use of international tests as a governing device, with politicians and policy-makers in some jurisdictions using PISA results to legitimate policies. Correspondence between test results and particular policies is not evident. Rather, it appears that, at least in some articles, governments have used the results to justify pre-existing policy trajectories. International test results can be used to legitimate policies as they can appear as an independent arbiter of standards – such uses are known as an externalisation device (Lawn & Grek, 2012). So, there is concern in this literature that the international test movement will produce policy convergence across nations, which could be detrimental to the expression of educational values within countries and to diversity generally. To some extent this has been observed, as there is an international trend in which more countries are introducing national (and international) tests, school and teacher accountability measures and so on. On the other hand, even when policies use the same nomenclature, they can often look very different in practice in each country. Going back to Hanson's (2000) argument that the signifier (test results in this case) can become more important than the signified (learning in this case), there is also an issue about the extent to which international tests receive attention and set the agenda. This could distract us from focusing upon potentially more important aspects of learning, with implications for resources and policy-making.

The production of these large-scale datasets has been a fruitful avenue for research for many and there is a great deal of secondary analysis of the data. Many of the articles conducting secondary analyses do not critique the underlying source of the data, although some do. Further, there are articles focused upon a critique of the systems and processes underlying the test data. We outline some of research relating to sampling, the psychometric models and test translation.

1.7 Using validity to examine relationships between assessment and learning

Paul Newton

Validity is the principal conceptual tool that assessment professionals use to interrogate assessment policy and practice. Throughout the 20th century, the concept became increasingly important and fundamental. Indeed, by the end of the century, many assessment professionals believed that it should signify, not simply the technical quality of an assessment procedure, but its social value too. Not everyone agreed with this

expansion of the concept, however. Vigorous debate ensued over the very nature of the concept; and this has escalated following the turn of the present century.

The escalation of this debate has cast new light upon many important issues that might otherwise have remained obscured. This is particularly true in relation to the complex relationships between assessment and learning; and this is why we have used validity as a lens through which to examine these relationships. We were particularly keen to uncover tensions that might exist between educational assessment practices and the cultivation of valuable learning.

Certain elements of the current debate enabled us to highlight technical challenges, such as the potential for extrapolating from heavily standardised assessment tasks to the rich, authentic, robust understandings that they are supposed to represent. Other elements allowed us to highlight social challenges, such as the need to acknowledge and accommodate the multiple perspectives and values that are embodied within and served by educational assessment systems. Perhaps most importantly, the current debate helped us to foreground the fact that both assessment and learning occur in localised, situated, dynamic and partially unique contexts, implying that our attempts to examine and evaluate their many relationships need somehow to embrace this complexity.

1.8 Doctoral work

Therese N. Hopfenbeck and Anna T. Steen-Utheim

Over the past 20 years, 29 doctoral theses were identified which have been published in the field of assessment in Norway. Seventeen of them were in the area of assessment and learning, eight on international tests, two on national tests and two on examinations. The surge in theses on international tests is in part due to government funding for research in this area. A much smaller number of theses had been conducted on the national tests and examinations. With the exception of one thesis, from the abstracts alone, it is not entirely clear the extent to which of the theses are focused upon AfL.

Identifying theses through electronic searches proved to be an unsuccessful strategy (only two were found). As some theses were gained through publications, there are copyright issues for some works. Other candidates may not wish their work to be published to avoid future copyright issues when publication in academic journals is sought. For the purposes of this review, we relied upon our academic network to advise on theses that had been published in this time period. There are, of course, limitations to this approach and it is possible that we have not identified all of them, especially those published a number of years ago. Also noteworthy was the fact that titles and abstracts often did not contain the keywords that would signal the content of the

thesis. This is likely to have a detrimental impact upon the dissemination of the doctoral work, as it will be difficult for researchers to identify the relevance of the work. We recommend that doctoral researchers produce abstracts of their thesis in Norwegian and English, ensuring that the focus of their research is reflected in the title and abstract, to produce more impact from these projects into which a lot of resource and effort has been invested.

The PhD dissertations in assessment and learning covered a variety of areas, such as school policy and school evaluation; portfolio assessment; international assessments; national tests; assessment in writing, dialogue and foreign languages; feedback, students' learning orientation; and identity and assessment. The dissertations also covered primary education (three), secondary education (seven), upper secondary education (two) and higher education (five) as well as themes across levels. Of the 29 studies, 15 used qualitative techniques such as interviews, observations and document analysis while ten used quantitative methods such as questionnaires and statistical analysis of items. Four of the dissertations used mixed methods approaches. One noticeable finding was that most of the quantitative analyses were conducted by candidates working with international datasets.

The Norwegian dissertations showed that there is more focus upon international tests than earlier, but fewer studies conducted on AfL or formative assessment. With the focus upon AfL programmes in Norway over the past few years, more research in this area might have been anticipated. Gamlem (2014; Table 3) and Eggen (2005; Table 3) are the only studies that specifically focus upon AfL or formative assessment. Further, there is a lack of intervention studies that measure learning outcomes and also a lack of randomised controlled designs.

Overall, the dissertations showed the development of considerable assessment research capacity in Norway, which bodes well for the future production of research and evaluation in this field within the country.

The state of the field: research in assessment and learning

Undoubtedly, assessment itself and its relationship with learning are important topics in education today. In this review, we have brought together aspects of the literature that are somewhat disconnected and highlighted how theoretical, philosophical, empirical issues and matters of policy and practice are playing out in those fields. In the process of conducting the review, and as others have pointed out previously, we saw that the field is fragmented and there is a lack of cumulative progress. Getting on top of the various, often somewhat technical and obtuse, literatures involved in this review is not an easy task, but in doing so, we found that there are a number of areas in which discussions are being conducted in parallel, without reference to what is being discussed elsewhere. One example of this would be the relatively infrequent cross-referencing between articles published on AfL and formative assessment.

To some extent, the difficulty and extensiveness of the literature helps explain this, as becoming an expert in one part of the field is a large investment. Additionally, there has been a proliferation of academic journals in education and it would be difficult for authors to find, let alone read, the burgeoning field of AfL articles. In part, the fragmentation is geographical, with authors in one part of the world citing research from their own part of the world, conducted within their educational and policy context. Equally, educational research encompasses broad schools of thought, with many different philosophical, theoretical and methodological stances coexisting in the literature. Some authors might reject the utility of parts of the field for paradigmatic reasons and therefore cross-referencing is unlikely. In keeping with Phillips (2014), we see quality in research as more than compliance with methods. Rather, quality is about the extent to which a compelling argument can be rendered that stands up to critical scrutiny. Thus, we see excellent work in the literature on psychometrics, assessment, validity, international tests and AfL that is based purely upon critical evaluation. Equally, a great deal of empirical work has been published which is of high quality – both qualitative and quantitative – on international tests.

Relatively few quantitative studies have been published relating to AfL. There has been a large volume of small-scale studies involving interview data with a small number of teachers and students in one or two schools. In itself, this is not problematic, as the insights gained from such small-scale studies can be very important more broadly. For example, Barbara Crossouard's (2011) study on formative assessment in conditions of social adversity was conducted in only two primary school classrooms in Glasgow, but moved the field forward. In that study, Crossouard grappled with the tension of the emancipatory function of education systems and the embracing of the sociocultural. Powerful knowledge was emphasised less in classrooms when teachers adopted a divergent approach to teaching and assessment which valued what pupils bring to school. So small-scale studies can produce rich, substantial descriptions that are compelling arguments for important findings. Many of the small-scale AfL studies reviewed for this report were also conducted by practitioners and thus the local knowledge gains and impact upon practice may well have been important. Some argue that relevance to practice is in itself a criterion for rigour (eg Gutiérrez & Penuel, 2014). However, all too often we found that these small-scale studies were not well informed about previous research. No doubt the timescale and financial pressures are at work, producing this pattern of widespread, small studies that do not build upon each other cumulatively.

In relation to the intersection of the fields of assessment and learning, we conclude that cognitive models of learning are currently most influential. If assessment is to continue to play such a large role in defining educational quality internationally, we suggest that assessment researchers and practitioners need to work more closely with substantive experts to elucidate, test and build better theoretical models to underpin the

assessments. We have also seen in this review that there is growing activity in the field relating sociocultural theory with assessment. As yet, there are some difficult and unresolved tensions that question assessment paradigms. This is an interesting field where creative work might find innovative solutions to these issues over the coming years. Some have characterised the field of educational research as an immature field (Wieman, 2014), lacking the underlying causal models that have been developed over centuries of research in the hard sciences such as physics or medicine. If we are to grow up, we need to be more connected as a field, which serves to underline the importance of thematic reviews that draw the field together and, in turn, of the Norwegian Knowledge Centre for Education.

1.9 Acknowledgements

We would like to thank all of those who gave us comments on earlier drafts of this report and supplied us with useful sources, including our advisory group, Professor David Andrich (University of Western Australia), Professor Eivind Elstad (University of Oslo), Dr Peter Gu (University of Wellington), Professor Trond Eiliv Hauge (University of Oslo), Professor Mary James (University of Cambridge), Professor Sølvi Lillejord (Norwegian Knowledge Centre for Education), Dr Erik Ruud (Norwegian Knowledge Centre for Education), Associate Professor Astrid Tolo (University of Bergen) and Dylan Wiliam (Institute of Education, University of London). We would also like to thank our research assistants, Anna Niedbala, Yasmine El Masri and Jeanne Ryan, and the project administrator at the Oxford University Centre for Educational Assessment (OUCEA), Joanne Hazell.

Additionally, we would like to thank the following people for their generosity in supporting us in the search for doctoral theses: Assistant Professor Toril Aagard (Telemark University College), Professor Eevi Beck (University of Oslo), Academic Librarian Hilde Kvein Bjerkholt, Humanities and Social Sciences Library, University of Oslo, Professor Stephen Dobson, (Hedmark University College), Professor Astrid Birgitte Eggen (University of Agder), Dosent Roar Engh (Buskerud and Vestfold University College), Associate Professor Henning Fjørtoft (Norwegian University for Science Technology), Assistant Professor Hege Hermansen (Oslo and Akershus University College of Applied Sciences), Professor Frøydis Hertzberg (University of Oslo), Professor Gunn Imsen (Norwegian University for Science Technology), Professor Rita Hvistendal (University of Oslo), Professor Lars Monsen (Lillehammer University College), Professor Jorunn Møller (University of Oslo), Researcher Dr Guri Nortvedt (University of Oslo), Researcher Rolf Vegar Olsen (University of Oslo), Researcher Astrid Roe (University of Oslo), Professor Atle Skaftun (University of Stavanger), Professor Kari Smith (University of Bergen), DPhil student Sverre Tveit (University of Oslo) and Professor Line Wittek (Buskerud and Vestfold University College).

The sculpture on the cover of this report depicts an oral doctoral examination and is to be found on the Examinations School building of the University of Oxford. The photographer was Yasmine El Masri.

This research was sponsored by the Knowledge Centre for Education, Norwegian Research Council (case number 13/4697). In particular we would like to give credit to Director Sølvi Lillejord, whose idea it was to conduct this review.

Notwithstanding all of the support we have had, the authors take responsibility for any remaining mistakes or omissions.

Chapter 2 Why a state of the field review is needed

Research in education has been characterised as being too small-scale and non-cumulative (for a review and critique of these claims see Oancea, 2005). Commentators have pointed to the need for overarching reviews so that the field can build upon what has come before (James, 2012). As such, state of the field reviews in educational research in general will benefit the literature. More specifically, a state of the art review on 'assessment and learning' is necessary because purposes, mechanisms and conceptions of assessment are evolving and the relationship between assessment and learning is therefore also in a state of flux (eg Pellegrino et al., 2001; Madaus et al., 2009, Pellegrino et al., 2012).

A prevailing narrative of our time is that countries now compete in terms of human capital as well as manufacturing and access to natural resources (Powell & Snellman, 2004). In this 'knowledge economy', education has assumed an even greater prominence than it had in the past (Schleicher, 2006). National and international assessments have become central to quality assurance and to improving education and are therefore crucial in policy terms as levers and controls of education systems (Higham & Yeomans, 2010).

High-level policy is far removed from classroom practice where the learning takes place and policy is not implemented wholesale in systems – it is shaped and formulated in the process (Moran et al., 2008; Ball et al., 2012). Notwithstanding this, the literature on washback indicates that assessments have come to define the taught curriculum to a larger extent than in the past (eg Darling-Hammond & Rustique-Forrester, 2005). At all levels – governments, schools, teachers and students – there is a greater desire to cultivate the necessary knowledge and skills to attain valued qualifications and to compete in the knowledge economy (Brown et al., 2011).

In this review, we look at two key developments in the literature on assessment this century: international testing and AfL. We look at not only those fields, but also how they are connected theoretically, empirically and conceptually with theories of learning. Different learning theories have distinctive positions on the definition of learning itself, the process of learning and therefore implications for its assessment. Theoretical work on the relationship between assessment and learning is not in abundance, but some authors have tried to establish the connections with learning theory in broad overviews (Shepard, 2000; Elwood 2006; James et al 2007). Additionally, there are different assessment paradigms within the field of assessment and psychometrics, which bring a variety of positions regarding how learning can be defined and measured (Behrens et al., 2012; Baird & Black, 2013).

Theoretical work on validity is concerned with how to conceptualise and how to evaluate the technical quality and social value of assessment. Conceptions of validity are

variable, with different theorists adopting different positions, but all now agree that broader issues of social value need to be embraced in any comprehensive evaluation of assessment procedure. Any review of the relationship between assessment and learning must therefore be situated in the context of the recent literature on validity.

Countries have increasingly implemented national testing programmes in response to this (eg Elstad et al., 2009) and we have also seen a rise in supranational tests this century (such as PISA or TIMSS), which influence the discourse on assessment and learning (Ozga, 2012). A key question arising concerns the impacts of national and international tests upon enacted curricula and upon learning. How do such assessments affect conceptions of learning? How do they influence what counts as powerful, valuable knowledge?

Alongside the above developments, we have also witnessed the rise of AfL (Black & Wiliam, 1998a) internationally. This approach to assessment focuses upon classroom practice and typically involves different assessment technologies to the national and supranational tests. Again there are questions to be answered about the desired and actual impact of this form of assessment upon learning and how this is distinguished from other forms of assessment.

These two developments – international testing and AfL – are very different in nature. They are different assessment technologies in terms of their purposes, practices and impacts. International testing is essentially designed to measure educational systems and its connection with learning practices is less direct than that of other assessments. However, its impact upon the definition of learning at governmental levels is increasingly evident. AfL is geared towards classroom practices and improvement, not just the measurement of learning. This form of assessment is situated in the teaching and learning environment. Contrasting these two distinctive developments which have had a large impact upon the literature in terms of their relationship with learning is an important addition to the field.

2.1 Why this review is important for Norway

In his 2010 keynote address to the Association for Educational Assessment – Europe conference, Petter Skarheim, the Director of the Norwegian Directorate of Education and Training, noted the following challenges for the Norwegian education system:

- Norway has a short history of research on assessment and learning;
- There is a lack of national data on learning outcomes – Norway has to rely on international datasets; and
- There is an overall lack of an assessment culture, competence and practice in schools and in teacher education (Skarheim, 2010).

Historically, Norwegian education research has not focused greatly upon learning outcomes such as achievement scores on tests. Instead, the main work on assessment in Norway in the 1980s and 1990s was on school evaluation, with a bottom-up approach, where municipalities and schools defined their goals for their local schools (Haug & Monsen, 2002; Moe, 2002). National tests were seen as alien in the Norwegian education system (Hertzberg, 2008), which has been heavily influenced by broad goals such as equity, inclusion and democratic values (Telhaug, 2006, 2007; Welle-Strand & Tjeldvold, 2002) – issues also discussed in a special issue of *Norsk Pedagogisk Tidsskrift*, in 2007 (2) see Smith (2007) and Eggen (2007).

The 2011 OECD Reviews of Evaluation and Assessment in Education in Norway (Nusche et al., 2011) observed that a national quality assessment system (NKVS) was set up in 2004 to provide a range of data and tools to help schools and education authorities to evaluate the performance of the system and use this information in their strategic planning. The system includes national student assessments, surveys, a web-based portal for schools and other tools and guidance to support local-level evaluation. The authors credited Norway for providing this toolkit for accountability and improvement and indicated that the country's top priorities for evaluation and assessment were to:

- Clarify learning goals and quality criteria to guide assessment and evaluation;
- Complete the evaluation and assessment framework and make it coherent;
- Further strengthen competencies for evaluation and assessment among teachers, school leaders and school owners.

(p9–10)

In a recent study, the previous Minister of Education, Kristin Clemet, described the lack of data on learning and assessment as one of the main challenges for the Ministry, when she first came into office in 2001, just months before the first PISA release. The Ministry had a lot of data about structures, resources, and economics in schools in Norway, but when she asked questions such as 'How do we know that students learn what they are supposed to learn in schools?' there was less knowledge and documentation available, including on how teachers assessed their students (Hopfenbeck et al., 2013).

As in many other countries (eg Germany – Ertl, 2006), the release of PISA in 2001, where Norway performed less well than the average of the OECD countries, came as a big shock to Norwegians, and resulted in increased media attention for each cycle as well as policy discussions around who was to blame for the results (Elstad, 2012). PISA was followed by the introduction of the National Quality Assessment System (NQAS), which has increased the focus on students' performance in terms of educational outcomes and can be described as a shift in the Norwegian educational policy from the

use of input-oriented policy instruments towards a more output-oriented policy (Skedsmo, 2011). NQAS includes national tests in reading (Norwegian) and mathematics in Years 4 and 10, and English reading in Year 10, as well as diagnostic tests, mapping tests, international tests, and student and parent surveys.

Municipalities are expected to use the system to monitor and develop their schools. Evaluation of the municipalities suggests that they differ in the way they use data and information from the NQAS to improve the quality of their schools, and that there is a real need for capacity building at the municipal level when it comes to knowledge of how to use data for quality development (Rambøll, 2013). It has also been found that not all school leaders have the capacity to use and interpret the results in the new national system (Aasen et al., 2012). In other words, there is still a demand for greater assessment literacy at all levels in Norway, as well as knowledge of what counts as learning.

To summarise, a state of the field review of 'assessment and learning' is needed because the field of educational research needs to be more cumulative, building upon knowledge gained in previous projects. This particular review is necessary because the relationship between assessment and learning has become increasingly central to education worldwide, with the growing emphasis upon measuring performance in the global knowledge economy.

How assessment defines valuable learning according to different theories and by different assessment technologies needs to be explored more fully. This is our overarching and cross-cutting theme that runs through our review. Two major areas of development in the field this century are then reviewed, for contributions to the literature on assessment and learning, and in relation to the overarching theme.

Chapter 3 Method

Jesson et al. (2011) indicated that approaches to reviewing the literature sit along a continuum on four dimensions. This review is narrative in style, but we adopted a structured approach with a rigorous method and we specify the analysis to some extent (Figure 1). A narrative review may address specific methodologies or studies, be thematic and conceptual in nature, introduce the most recent research on a subject, express the position of acknowledged experts or scope the field to set future research agendas. Our review is at the intersection of the fields of assessment and learning and we explore what has been written on this topic, as well as making our own observations. As the review is so broad, it is not a narrative review that digs deeply into specific studies or methodologies – some topics we mention only in passing. The review is more conceptual in nature, making connections and distinctions across the field. It is a state of the art review, so we focus upon research conducted this century, but address key articles in the field where necessary.

Figure 1 Continuum of literature review approaches

Narrative review	⇦✓----------⇨	Systematic review
Variety of styles	⇦----------✓--------⇨	**Structured approach**
No defined method	⇦----------✓-----⇨	**Rigorous method**
No specified analysis	⇦----------✓----------⇨	Synthesis, meta-analysis

(adapted from Jesson et al., 2011, p11)

To underpin the writing with a rigorous method, we conducted systematic reviews of the literature on international tests and AfL. We outline the broad approach taken for each of those below (3.2) and go into more detail in sections 5.1 and 7.1. The chapters on valuable learning (Chapter 4), views of learning in the psychometrics tradition (Chapter 6) and validity (Chapter 8) were written purely as narrative critical evaluation reviews. The structured reviews on international tests and formative assessment informed the critical narrative reviews that were written on those topics. This approach was used to ensure that the review covered the existing field and did not miss important work on international tests and AfL. Next, we explain why we adopted a narrative critical evaluation method.

3.1 Narrative critical evaluation review

A number of different approaches to reviewing the literature were available to us. Although there are advantages to systematic review methods (explicit procedures, extensive search, clear quality criteria, reduction in reviewer judgement, combination of findings), we largely conducted a narrative review. This avoids the pitfalls of systematic reviews, which adhere to a positivist epistemology (Hammersley, 2001; MacLure, 2005) and tend to incorporate a hierarchy of credibility in research designs (from anecdotal to systematic review and meta-analysis). Whilst we accept that certain designs are advantageous in research that seeks to establish causality, many of the key articles in the field of assessment do not conform to this hierarchy and would be excluded from the review in a systematic approach. Thus, we included critical evaluation articles, empirical work (quantitative and qualitative), randomised controlled trials, systematic reviews and meta-analyses, so long as they made an important contribution to the field. Additionally, although we do not include the grey literature generally, we include book chapters as well as papers published in peer-reviewed journals and reports of national or international standing (eg Dumont et al., 2010). It is our observation that some of the central papers in the field, particularly in Europe, are published in books. We sought to avoid the problem of systematic reviews in which too many articles are ruled out of the process (see MacLure, 2005, p401).

Further, we did not seek to remove reviewer judgement from our review. The authors' collective judgements and interpretations are essential to the quality of the review. Hammersley (2001, p549) wrote that the reviewer needs to 'draw on her tacit knowledge, derived from experience, and to think about the substantive and methodological issues, not just to apply replicable procedures'. This will be a critical evaluation of the key developments in the literature, focusing upon the importance of articles for their conceptual, as well as empirical, contributions to the literature. Our aim was to produce a high-quality review that embraces the following features; 'intertextual connectivity, critique, interest, expertise, independence, tacit knowledge, chance encounters with new ideas, and dialogic interactions between researcher, "literature" and "data"' (MacLure, 2005, p394). Next, we outline the general approach to the systematic reviews.

3.2 Systematic review method

A systematic review method was applied to the literature on AfL and on international tests. Specific details of the method not included here, such as the search terms, are given in sections 7.1 and 7.1. Systematic searches were conducted using the Oxford Bodleian Library's SOLO system and Web of Science. References were saved and categorised on EndNote software. Additionally, the EndNote online search function was used to search Academic Search Premier, ERIC, and Zetoc. Cross-referenced or related

articles that appeared in online databases when the citing article was retrieved (eg through Science Direct's Related Articles feature) were included. A search for articles citing each retrieved reference was conducted using Google Scholar. Where possible, PDFs were retrieved and stored in the database.

Duplicate references were removed using the 'Find Duplicates' function in EndNote. Any duplicates not detected by this tool were removed manually. This could occur due to, for example, spelling or capitalisation differences. References that did not contain one of the search terms within the body of the article (as determined by a search of the text contents) were removed. For example, several articles cited discussions of one of the assessments in question, but did not directly discuss them. Articles that were irrelevant to the topic (such as biomedical articles using the same acronyms and several unrelated articles with affiliations to the University of Pisa) were removed from the EndNote library as well. Articles that only contained the search terms in the references section were removed.

The Update References function in EndNote was used to ensure that as much accurate information as possible was stored for each reference. Based on the article titles and abstracts, keywords for each reference were updated manually with the topic, context, country, test(s) discussed, and methods. Only peer-reviewed articles were included in the systematic searches. Articles were not ruled out of the search on the basis of language, but where possible, the English version of the abstract was included in the database. Thus, the databases included articles in English, French, German, Norwegian, Spanish, Swedish and Turkish. Categorisation schemes were devised and applied to each database using the information in the abstracts and PDFs (where available). Over 850 articles were categorised in the international test database and over 900 in the AfL database.

3.3 Limitations

The review was conducted in a short time period, which does not allow for a detailed scrutiny of all of the papers in the field. We sought to counter this limitation by drawing upon our extensive knowledge of the literature to date, our previous projects and discussions around the state of the field. Our Senior Project Advisers (see p172) gave us feedback about the extent to which we give a fair representation of important works in the literature.

We now turn to the narrative review, beginning with an explanation of the centrality of the field of assessment in education and why it has risen to such prominence.

Chapter 4 Valuable learning

Jo-Anne Baird
OUCEA

Learning is ubiquitous, but education is about cultural transmission and the cultivation of *valuable* forms of learning in societies. In *Education for All. The Future of Education and Training for 14–19-Year-Olds*, Pring et al. (2009, p13) described the concerns of education as

> …introducing young people to a form of life which is distinctively human, which enables them to be and to feel fulfilled, which equips them to live independent economic lives, and which enables them to participate positively within the wider community.

We do not mean to imply that only through formal education can valuable learning take place. Much important learning happens outside the classroom that affects people's lives, livelihoods and communities more broadly. Rather, we intend to recognise that a purpose of education is to convey what has been deemed important for people to know and be able to do. What is important to learn in a particular time and place may not be deemed so valuable in another. For example, use of computers was not part of the school curriculum 100 years ago because they had not yet been invented. Curricula are context-dependent.

Educational assessment has increasingly come to define curricula (Au, 2007; Madaus et al., 2009), partly because of its uses as levers and controls by governments (Higham & Yeomans, 2010). Qualifications are also important currencies for individuals' educational and occupational progression. Assessments therefore set the agenda to a larger extent than they did in the past and they provide or withhold the keys to better life chances. Consequently, assessments define what counts as valuable learning and assign credit accordingly.

Meritocracy is, of course, an imperfect description of societies. As such, the relationships between the credits assigned through assessment results and life chances are less than perfectly correlated. Other factors, such as the social capital a person has through personal networks, intervene. Socioeconomic status at birth can convey a wide range of benefits or disadvantages. Additionally, Brown et al. (2011) pointed out that qualifications are not a straightforward ticket to success because global forces have produced highly qualified workforces in the developing countries who are prepared to work for less than those in the west. So we acknowledge that good assessment results are not a one-way ticket to success. However, they are frequently part of the gate-

keeping armory for jobs and educational places and as such define what counts as valuable learning and allocate points accordingly. Often the details of how students must demonstrate their knowledge and skills matter for the ways in which they learn. What counts as credit-worthy on an assessment rubric affects what is learned. More than this, Hanson (1993, 2000) argued that assessments actively create and shape people's identities. Conceptual clarity about what intelligence means is hard to come by, but we have intelligence tests that are used as a tool to measure and thus define what counts as smart. Once classified, all kinds of gates can be opened or closed to us on the basis of the test scores.

It would be remiss to pretend that even this is the whole story as there are other levels at which assessment scores are put to use. Researchers have consistently pointed out that the use of educational assessment in modern society makes its impact felt well beyond even the curriculum, classroom practice and learning. Assessment has been an integral part of neoliberal approaches to education (Gergen & Dixon-Román, 2013; Ball, 2008). Such an approach sees many aspects of society as a marketplace. In the educational sphere, professionals' outputs must be measured, targets set and efficiency determined. So much of this is done *to* individuals and organisations by central administration that there is alienation and disjuncture from the real content of learning, as well as from the context of assessment. Learning is decontextualised and reduced to a number.

Lyotard's (1984) notion of 'performativity' is key here. He referred to the 'mercantalization' (p51) of knowledge, which we also see in discussions of the knowledge economy. Performativity is the maximisation of efficiency in this mercantile system of knowledge. Questions about truth, Lyotard argued, are replaced by notions of the utility of knowledge. As Torrance (2000, p179) pointed out, this effectively replaces the end with the means. So, working within such a system, people must look to the currency of knowledge and to demonstrating and maximising their productivity within that system. These systems, or 'signifiers' of knowledge, in Hanson's terms, come to be seen as more important than the knowledge itself. Current focus upon test results, school performance and national performances bears witness to this argument. At the levels of government, schools, teachers and students, the numbers produced by tests matter in our social systems and therefore people strive to attain them. As Hanson (2000) predicted, the signifier is often seen as more important than the signified, as the discussion has shifted to the assessment results, rather than the learning or the curriculum. Discussions are often generic with respect to subject discipline and the content of the tests or the ways in which the results are produced is neglected.

Relatedly, Michael Power (1999, 2000) wrote about the 'audit society' and the rise of new public management, with its increased demands for accountability and transparency. Such systems actually foster distrust (O'Neill, 2005) and result in disengagement. They are centralising, standardising forces which serve to undermine

local values and, ultimately, human capacity (McNeil, 1988; Shepard, 2000). For example, few teachers in England are now confident in their competence to write a curriculum, since it has largely been centrally controlled for the past two decades. Consultations on proposed curriculum changes in England have often gone unchallenged by teachers. This is in marked contrast to the days before the national curriculum was introduced.

Testing is not a neutral activity. The content and form of testing, as well as its uses, are defined by those in powerful, institutional positions. Consequently, national testing serves to conform with, recreate and justify existing societal hierarchies. Challenges to these values have to take on not only the testing edifice, which is formidable, but the ideological neoliberalism which it scaffolds. Given the powerful position of educational assessment in defining valuable learning, it is important to look at how learning is conceived of in the theory and practice of assessment.

4.1 Theories of learning and educational assessment

Underlying theories or approaches to defining learning are not always explicit or obvious in the literature in this field. Developments and tensions in the field of learning theory (eg for an overview see de Corte, 2010) have coexisted with theoretical and technical advances in assessment and psychometrics. James (2006) provided an overview of the introduction and acceptance of different assessment formats (eg examinations, coursework and portfolios) over time in public examinations in England. She outlined the links with behaviourist, cognitive and sociocultural theories of learning, but accepted that there was not a one-to-one correspondence between these theories and assessment formats.

In this review, we categorise learning theories following James (2006). We recognise that assessment practices may often implicitly draw from a combination of theories. For example, a multiple-choice test may reflect a behaviourist perspective while accompanying open-ended questions may suggest a more constructivist approach. We are also aware that examinations such as the Chinese civil service selection examinations, the model for much that was developed subsequently, pre-date these learning theories by many centuries. In this sense many assessment practices have not been theory-driven, even though there may have been implicit assumptions about learning and the role of assessment in this.

For the first half of the 20th century, a behaviourist theory of learning had precedence. In this conceptualisation of learning, mental states are not of much interest and the main focus is upon how people behave in relation to their environments. Behaviourism was closely aligned with positivist (Suppe, 1984), scientific principles and drew much of its credibility from mathematically-based theories which were predictive of behaviour,

especially in animals. For example, the gradual extinction of learned behaviour when it is only partially reinforced could be plotted (Skinner, 1938). Although behaviourism's reliance upon stimulus-response patterns could seem too rudimentary to explain how we come to learn complex matters such as Pythagoras' theorem, other scientific theories with seemingly simple bases were also known to be very powerful for prediction. Evolution, for example, had produced very complex organisms from a simple selection principle: survival of the fittest. Small steps towards the desired behaviour were the key to shaping performance so that learning occurred in behaviourism.

James (2006) argued that traditional testing reflects this behaviourist model of learning and others have similarly expressed this view (Black, 1999, p120). The work of Skinner (1938, 1950) and Bandura (1971) were very influential in the shaping of this field of research. Nowadays we are not always so impressed with the reproduction of facts without a demonstration of understanding. Bloom's famous (1956) taxonomy has been very influential in the design of educational assessments. In this taxonomy, a hierarchy of cognitive demand in educational objectives were outlined, with knowledge being the lowest level and evaluation being the highest (Figure 2). This was later revised (Anderson et al., 2001), to incorporate two dimensions, relating to what we know and can do (Figure 3).

Figure 2 Bloom's original taxonomy (1956)

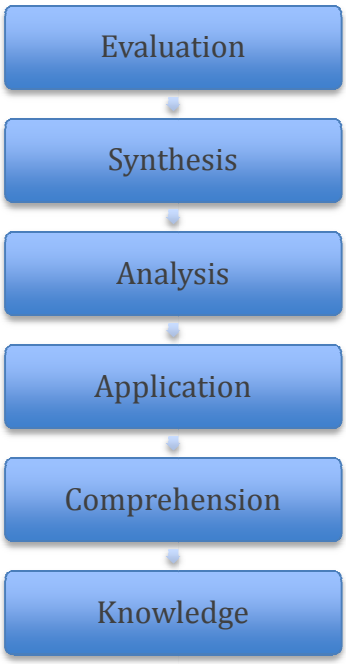
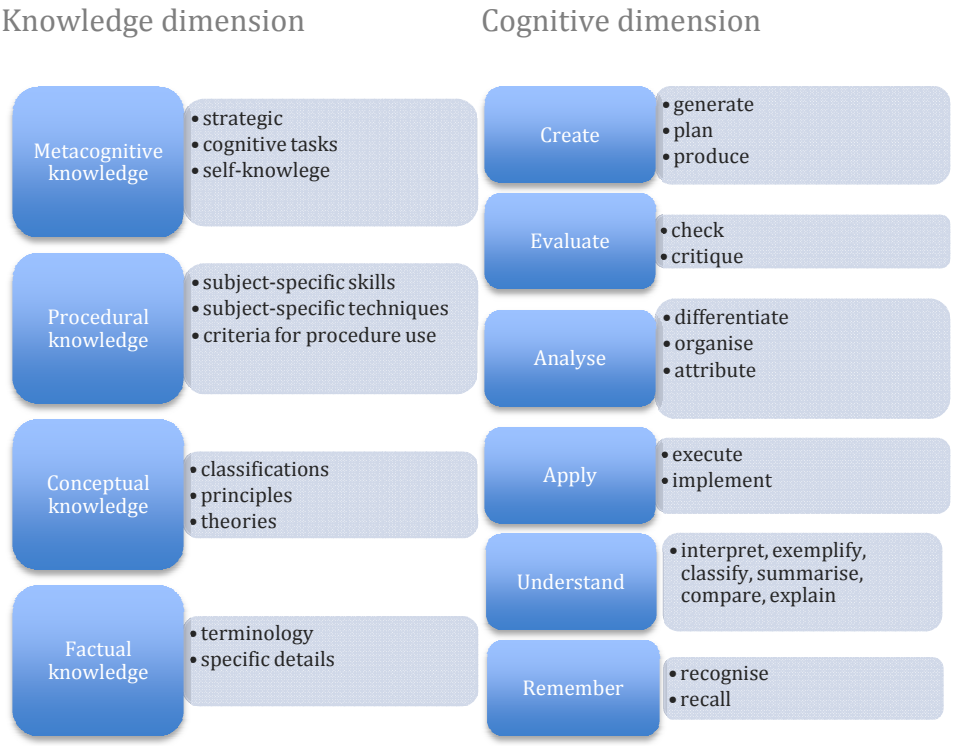


Figure 3 Bloom's revised taxonomy (Anderson et al., 2001)



Cognitive constructivist views of learning became more accepted in the 1960s and these raised issues of the thinking skills being brought to bear in learning. Mental processes, their organisation and their connections with previous knowledge and experiences became important, as did outlining the progression from novice to expert and the many slips people make along the way. The work of Piaget (eg 1967) was important to the development of cognitive psychology, but the fact that much of the developments were based upon a computer model of the mind (eg Boden, 1988; Searle, 1990, Anderson et al., 1997) is often overlooked in modern accounts (eg Black, 1999, p120; Pellegrino et al., 2001; James, 2006).[2] Cognitive views of learning are currently the dominant paradigm in assessment (Haladyna & Rodriguez, 2013, p29).

James (2006) argued that partial credit for a good thinking process (even if the end result was wrong) became a more frequently used scoring mechanism and extended responses were considered useful in allowing students the opportunity to explain their thinking and evaluation of materials that they had been exposed to. Note that partial credit and essays were not brand new assessment formats, invented in the face of cognitive constructivist theories of learning. Instead, James' argument is that they were more accepted and prominent due to the rise of cognitive views of learning.

A third view of learning stems from sociocultural theory, which has been heavily influenced by Vygotsky (1978). Mary James (2006, p56) pointed out that this tradition

[2] Freud's metaphor for the mind was the steam engine.

can be traced back to William James (1890). In sociocultural theory, learning is not seen as something acquired by the learner from the teacher, but as something that is jointly created between the learner and their social environment. According to this theory, students can create new knowledge and therefore assessment needs to capture this by utilising a variety of techniques that are closely tied to the learning situation (eg learning portfolios or coursework; Elwood, 2006). In a sociocultural theory, learning (even thought) is not something that exists only in the head. The mind is situated in social action and learning is something that arises from a joint activity of agreeing to call something learning (see Cobb, 1999). Learning is a process of enculturation and is mediated by artefacts and language. For example, Korean students outperform those in many other countries in tests of mathematics (eg Mullis et al., 2012). In part, this has been ascribed to the nature of the language, as number words are more explicit than in English, for example. The number 11 is expressed as ten and one. Abacuses are used, which reinforce the decimal number structure, and textbooks use different colours for the tens and hundreds numerals to help mediate their meaning for children (Bernstein et al, 1994, p58).

Torrance and Pryor (1998), James (2006) and Shepard (2000) argued that AfL is more closely related to this sociocultural theory of learning than other assessment practices (see section 5.3.3 of this review). This approach can involve not only sharing of assessment criteria, but the negotiation of these criteria between students and the tutor (eg Elwood & Klenowski, 2002). Further, it involves embracing the social interaction between students and between teachers and students (Black & Wiliam, 1998a, 1998b; Shepard, 2000).

Some authors have explored the ways in which sociocultural theory can be applied to assessment design (eg Moss et al., 2008; Pryor & Crossouard, 2008; Crossouard, 2011; James & Lewis, 2012). A variety of assessment techniques are explored in this literature – portfolios, questionnaires, diaries, communication with parents, peer assessments and so on. As discussed above, techniques associated with AfL are often related to sociocultural theory. AfL convincingly embraces the social aspects of learning, but only to some extent. Learners' views of the social are not well dealt with in this approach currently, as research in this area tends to be written from the teachers' perspectives. Further, as Elwood (2006) pointed out, AfL does not deal well with the cultural aspects of learning. Large-scale, 'standardised' assessments have the benefit of being the same for everyone, which has qualities of fairness. In effect, standardised assessments and AfL both ignore the cultural aspects of learning and the influences of societal structures. Application of theory to practice is challenging and it is encouraging to see these attempts to draw sociocultural theory into assessment practices. However, it is not yet clear that this has been successful. Whether any of these practices are grounded in sociocultural thinking is questionable; certainly we could justify these practices using other learning theories, such as cognitive constructivist approaches. Neither do sociocultural approaches uniquely point to certain practices. As such, at present the use

of sociocultural theory is neither necessary nor sufficient to justify specific assessment practices.

Although the field of learning theories is often presented as progressive, culminating in sociocultural theories, in truth these theories have coexisted for a long time (indeed, James (2006) recognised this). Not only have the theories of learning been in use and development contemporaneously, but the causal links between these theories and assessment theories and practices are difficult to unpick. Assessment is a field of practice that is only loosely connected with our underlying views of learning. Particular practices may be influenced by certain theoretical approaches, but these links are hard to firmly establish. For example, we often see AfL described as a progressive, cognitive, even sociocultural approach to assessment, but some of the practices are entirely in keeping with behaviourist theory. In a 1989 letter to *Science*, B F Skinner, a chief proponent of behaviourism, wrote that

> Good instruction demands two things: students must be told immediately whether what they do is right or wrong and, when right, they must be directed to the step to be taken next.[3]

The notion of feedback and scaffolding of the way ahead are key features of the AfL approach. So, assessment practitioners might agree at a level of practice, whilst disagreeing at a theoretical level and *vice versa*.

Some authors are explicit about their beliefs relating to learning theory and philosophy, but implicit historical legacies in practice also shape our assessment systems. Before turning to theories arising from psychometrics, let us consider some of the dilemmas arising from this consideration of learning theories.

1. *Functionalism versus traits* – Are we interested in behaviour (performance) or mental attributes?
2. *Cognitive versus sociocultural* – Do we assess the individual, or take account of the social and cultural nature of performance? These issues have significant challenges for the field of assessment, as it currently stands. How do we assess the individual, working with artefacts, within groups and with wider forms of cultural capital?
3. *Positivism versus interpretivism* – To what extent do we see assessment as an objective, scientific process, or as a subjective, social process?
4. *Realism versus postmodernism* – Are the phenomena that we are assessing real properties of individuals, or are they social constructions?

We return to the dilemmas above when discussing the psychometric tradition of assessment.

[3] We presume that Skinner meant that students be directed to the next step even if they were wrong, although he did not write that.

Different views about what counts as valuable learning coexist and create tension – which knowledge and skills are valued in each subject is debated. Often the underlying dilemmas are represented implicitly in these debates. James (2006) tackled national assessments, as distinct from the field of psychometrics. Gipps (1994) indicated that assessment differed from psychometrics because it

- did not see learning as a fixed property of the individual, but as something malleable
- was criterion- rather than norm-referenced
- focused more upon validity in assessment design (whereas much of psychometrics perhaps unnecessarily erred on the side of reliability)
- relied upon formats that assess higher-order thinking in depth
- was designed to produce the best performances from individuals with clearly presented, relevant, concrete tasks that were not overly anxiety-provoking.

Baird and Black (2013) argued that psychometrics can seem like an answer to somebody else's problems if the educational context of assessment is prioritised over other considerations. They pointed out that psychometric techniques are not well-suited to changing curricula, transparent assessment criteria, connections between questions, qualification focus rather than item focus, and multiple dimensions in tests, amongst other things. They argued that to move test theory forward, psychometrics must grapple with educational purposes to a larger extent. In fairness, psychometrics has influenced educational practices in some countries, such as the US (Moss, Pullin, Gee & Haertel, 2005), more than others. However, the authentic assessment movement goes beyond a critique of psychometrics to a redesign of assessments more generally in the name of validity. Related to the arguments regarding generalisation (see section 8.6), Desmond Nuttall put it as follows:

> With the benefit of hindsight, it seems strange that so much effort should have been put into the development and validation of general paper-and-pencil tests, when everything points to their artificiality, their remoteness from the nature of any normal job and the unelaborative conditions of administration. It seems likely that conditions of utility and reliability have prevailed over considerations of validity. The signs are now that validity is claiming its rightful pre-eminent position and that the careful specification of the universe of generalization is helping to stimulate improved conditions of assessment and more thought about evoking an individual's best performance.
>
> (Nuttall, 1987, p116)

Arguments for authentic assessment[4] were underpinned by philosophical justifications related to postmodernism (Gipps, 1994; Lewy, 1996), which recognise that assessments are socially constructed and their results interpreted through the observers' lenses. Pamela Moss' work (discussed in section 8.5) critiques standardised testing using postmodern philosophy and has often drawn upon classroom-based assessments, considered to be more authentic than standardised tests. 'Assessments' in this literature have incorporated standard examination techniques as well as more authentic performance formats: it is a broad term. The postmodern turn to assessment rather than psychometrics raises many legitimate questions, but our reading of the literature is that assessment and psychometrics research is still generally grounded in modernist practices (Mislevy, 1997), although there are calls for different ways of thinking in the future.

In this chapter, we have discussed the term 'assessment'. No formal, mathematical model was presented for this term because many models can be used alongside assessments of different kinds. Indeed, some assessments may be purely qualitative. Large-scale educational assessment and psychometrics usually depend upon the use of statistical techniques. In the next section, we turn to the literature on AfL and investigate how learning theory has been used in that area. We return to the views of learning inherent in psychometric approaches to assessment in Chapter 6.

[4] What counts as 'authentic' assessment has itself been a matter of debate. For example, see Gulikers et al. (2004).

Chapter 5 Assessment for learning and formative assessment

Gordon Stobart and Therese N. Hopfenbeck
Institute of Education, University of London, and OUCEA

AfL has been in the ascendency in research and in practice internationally this century. Our interest here is to trace the relationships between assessment and learning theory, as well as assessment's influences upon learning practice and the evidence for this. As discussed below, AfL incorporates a range of techniques, including formative assessment. While the term *formative assessment* emerged from a behaviourist theoretical framework it has progressively broadened to incorporate elements that align with other theoretical perspectives which often pre-dated behaviourism. As a consequence current formulations of formative assessment draw on a variety of theoretical strands, which are reflected in subtle differences in definitions and in different emphases in the implementation of formative assessment. This diversity has led to a critique of formative assessment which argues that it is now a cluster of approaches, making it difficult to separate the contributions of the different elements and therefore to evaluate their effects.

The position taken here is that current understandings of formative assessment can be aligned with a range of learning theories and that each will be reflected in differing formulations and practices. One key area is the relative role of the teacher and the learner. In the original formulation the feedback to, and response of, the teacher was in the foreground, whereas more recent theorising has often focused on the development of self-regulated learners and on broader social interaction. Paradoxically, this recent theorising draws, often implicitly, on learning theories developed before the rise of behaviourist learning theories in the US.

Another key issue is the extent to which formative assessment involves a set of generic teaching and learning skills and practices or has domain-specific features that will lead to differing understandings and practices in different subjects. The approach here is to present a broad overview of what is understood by formative assessment and then to trace some of the theoretical underpinnings and how they support the broadening of what are considered formative assessment practices. This will be followed by a summary of recent critiques of formative assessment and its claims. We then go on to look at the challenges of implementing AfL, as AfL cannot fulfil its promise if it is not implemented in the classroom.

5.1 A systematic search for AfL and formative assessment literature

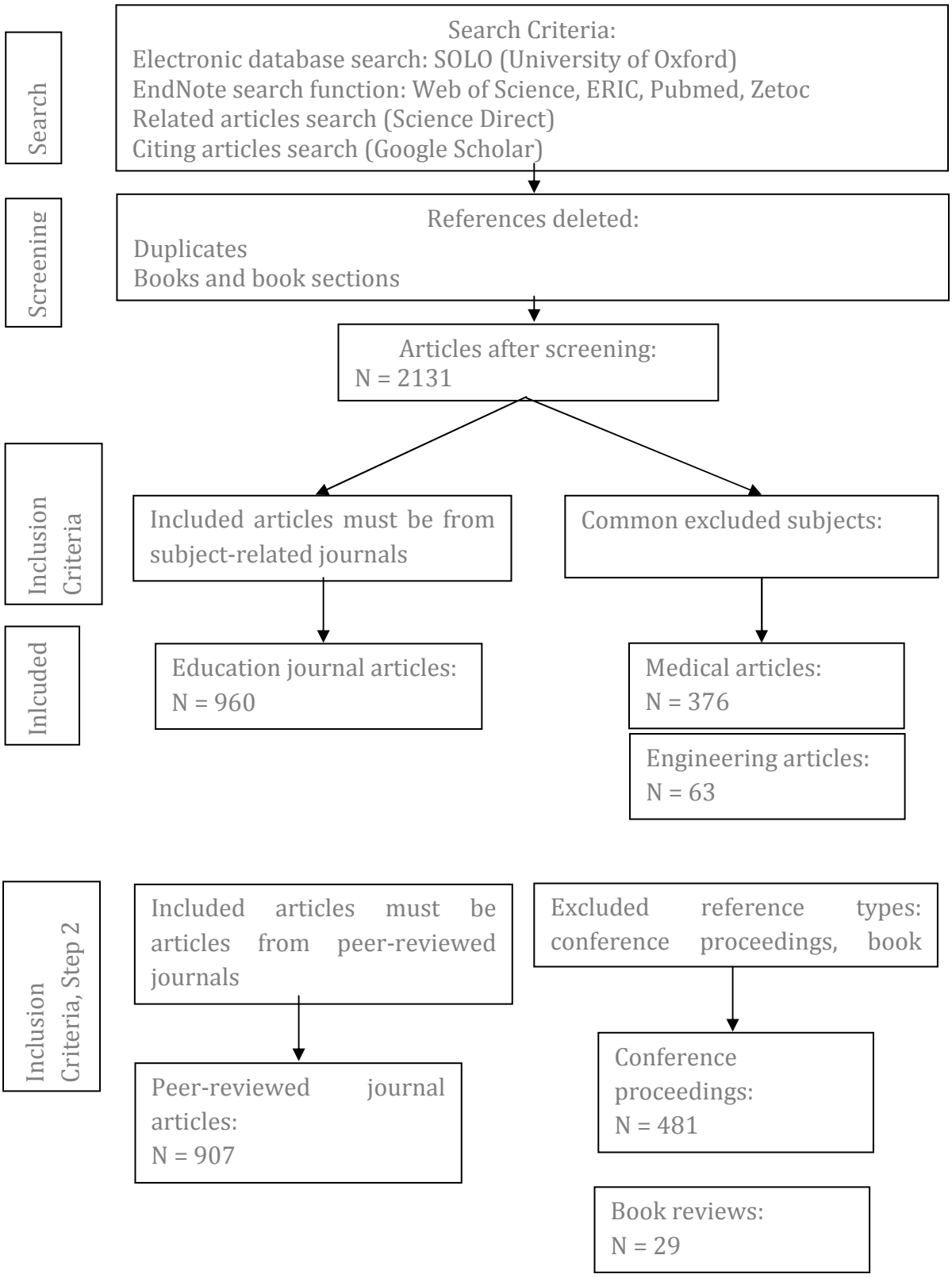
Our general method for conducting the systematic reviews was described in section 3.2. Here we describe the systematic search details specific to AfL. The following search terms were used: 'assessment for learning' and 'formative assessment'. A total of 2,248 references were retrieved and stored in an EndNote library (see figure 4).
Articles were stored in groups according to the search terms through which they were retrieved ('formative assessment' or 'assessment for learning'). We used Black and Wiliam (1998a) as a starting point for the AfL search and removed articles before 1998, leaving a total of 2,131 articles in the Endnote library.

Because a wide range of subject matter was included in the search results, including a significant number of medical assessment articles, references were filtered for articles from education and psychology journals. A list of all journals included in the library, as generated by EndNote's lists function, was analysed and 394 journals in the field of education and assessment were selected for inclusion. These references were sorted into peer-reviewed journal articles (907), book reviews (29), and conference proceedings (481). Articles from the most common non-education subjects—medicine (376) and engineering (63)—were excluded. For the 907 articles in the field of education, we read the abstracts and categorised them according to whether they were small or large-scale studies, case studies, and theoretical or empirical.

For the present review we were particularly interested to see whether we could find any large-scale studies implementing AfL or whether any studies had been conducted since 1998 which had tried to measure any effect upon students' learning. Reading the 907 abstracts, only two reviews of AfL were found since the review by Black and Wiliam (1998a) that tried to measure the effect upon learning quantitatively. Another pattern observed is that very few articles have been conducted on AfL on a large scale, and only one article has used a randomised controlled design. Most articles fell into the category of (1) case studies (and often one or two schools, with fewer than ten participants – such as the study by Hollingworth, 2012 and Sato 2005), (2) theoretical articles (Black and Wiliam 2009, Sadler 2010) or (3) action research (Harrison 2013, McDowell 2008).

Very few abstracts reported how studies had been conducted and which methods had been applied. Most abstracts did not give any information on how many participated in the study or what methods of analysis had been carried out. Most empirical research studies focused upon issues such as students' perceptions of AfL, either their experience of feedback from teachers (Hounsell et al. 2008) or their beliefs about AfL. Very few reported anything about learning outcomes. We do not report the findings of this systematic search here. Rather, we used it to inform the narrative review which follows.

Figure 4 Systematic search: AfL



5.2 Formative assessment[5] and learning theories

> Without a theoretical model of the mediations through which an interactive situation influences cognition, and in particular the learning process, we can observe thousands of situations without being able to draw any conclusions.
>
> (Perrenoud, 1998, p95)

5.2.1 Origins

There is general agreement that the use of the term *formative evaluation* originated with Michael Scriven (1967) in relation to programme evaluation. This stemmed from a debate with Lee Cronbach, who was arguing for the use of evaluation for ongoing improvement rather than at the completion of a programme (Cronbach, 1964). Though recognising this role Scriven emphasised the importance of summative evaluation and suggested 'formative evaluators should ideally exclude themselves from the role of judge in the summative evaluation' (1967, p45), a position he later revised (see Carless, 2011, p28–30; Roos & Hamilton 2008, p7–8 for a fuller account).

5.2.2 Widening the theoretical base

While behaviourist and neo-behaviourist approaches to learning may have led to the concept of formative assessment (formative evaluation in American usage), which was embodied in *mastery learning* (Bloom, 1984) and *measurement-driven instruction* (Popham, 1987), other theoretical orientations have contributed to a broader understanding of formative assessment. Cognitive and constructivist approaches placed more emphasis on the role of the learner in making sense of what is being learned and on the importance of self-regulation of learning. Sociocultural approaches emphasised the importance of context and the role of social activity in creating understanding. These orientations, which overlap, encouraged the use of a wider range of classroom strategies as part of formative assessment.

In the following sections an account of the main contributions of some differing theoretical orientations is provided. Two cautions are needed here: firstly, that there are considerable overlaps between both the theories and practices – the differences may be in what is foregrounded – and secondly, that much of the published work on formative assessment has implicit rather than explicit theoretical bases.

---

[5] There are sometimes confusing differences in the use of the terms *assessment* and *evaluation*. In the UK and some European usage, *assessment* is used in relation to judgements about classroom and individual performance. In the US the term *evaluation* would be used for this. Conversely, *evaluation* in the UK is used in relation to judging schools or systems ('school evaluation') whereas *assessment* is used this way in the US. This confusion is highlighted by the term 'teacher assessment' – in UK usage this would be the teacher's assessment of students whereas in the US it would refer to the evaluation (UK) of the teacher's performance. This is compounded by other languages not distinguishing between the two concepts and just having one word (eg in Norway the word is *vurdering*). UK usage is adopted in this review.

5.2.3 A working definition of formative assessment

The plentiful definitions of formative assessment can be treated as 'variations on a theme'. These variations may well reflect the learning theories, either explicit or implicit, which underlie them. Gregory Cizek (2010, p8) has synthesised current formulations into ten characteristics:

1. Requires students to take responsibility for their own learning
2. Communicates clear, specific learning goals
3. Focuses on goals that represent valuable educational outcomes with applicability beyond the learning context
4. Identifies the student's current knowledge or skills and the necessary steps for reaching the desired goals
5. Requires developments of plans for attaining desired goals
6. Encourages students to self-monitor progress towards the learning goals.
7. Provides examples of learning goals including, when relevant, the specific grading criteria or rubrics that will be used to evaluate the student's work.
8. Provides frequent assessment, including peer and student self-assessment and assessment embedded within learning activities
9. Includes feedback that is non-evaluative [=non-judgemental], specific, timely and related to learning goals and provides opportunities for the student to revise and improve work products and deepen understandings
10. Promotes metacognition and reflection by students on their work.

This is a catch-all synthesis from which exponents typically select or emphasise particular characteristics, for example the role played by teachers (2, 4, 7, 9), by the learners themselves (1, 6, 10) and in-classroom interaction between them (3, 5, 8).

Susan Brookhart (2007) summarised the progressive broadening of the concept of formative assessment in four stages. In her developmental sequence, formative assessment provides information about

1. the learning process;
2. the learning process *that teachers can use for instructional decisions*;
3. the learning process that teachers can use for instructional decisions *and students can use to improve performance*;
4. the learning process that teachers can use for instructional decisions and students can use to improve performance *which motivates students.*

These broadly align with the emphases of the different learning theories that have addressed formative assessment. The first two stages represent formative assessment's behaviourist origins. The third summarises the more cognitive/constructivist

contribution while the fourth incorporates both constructivist and sociocultural approaches.

5.2.4 Behaviourist and neo-behaviourist approaches to formative assessment

Scriven's use of formative evaluation was then adopted by Bloom et al. (1971) in relation to individualised teaching and learning, an approach then extended to group instruction (Bloom, 1984; see Wiliam, 2011, for a summary of these early approaches). For them, formative assessment involved 'the process of curriculum construction, teaching, and learning for the purpose of improving any of these three processes' (Bloom et al., 1971, p117). These developments were part of the 'mastery learning' movement which had grown out of the behaviourist learning theories commonly identified with Ivan Pavlov, Edward Thorndike, John Watson, B F Skinner and Robert Gagne. The central idea of the radical behaviourism associated with Pavlov, Watson and Skinner was that learning is a product of conditioning. In order to study this scientifically only observable stimulus and response could be studied – there was no place for invisible intervening variables such as the 'mind'. The key to learning was for the experimenter or teacher to manipulate the environment to create the desired responses, particularly through reinforcement and corrective feedback. Some of the core features of this approach remain central to other theoretical traditions, though the interpretation of them may have been modified. So, for example, Gagne's (1985) 'nine events' that take place within instruction are:

- Gaining Attention
- Informing Learners of the Objective
- Stimulating Recall of Prior Learning
- Presenting the Stimulus
- Providing Learning Guidance
- Eliciting Performance
- Providing Feedback
- Assessing Performance
- Enhancing Retention and Transfer

While the terminology may have been modified, the basic concepts of making the learning objectives clear, linking them to prior learning, and providing instruction and feedback are common to all theoretical approaches.

Current classroom practices that draw upon a behaviourist orientation are still influential, particularly in the United States. The widespread use of tests which publishers promote as 'formative' can be seen as a descendant of Bloom's 'test and remediate' approach. This encourages the idea that formative assessment is a product – a diagnostic test that assesses what is known (Bennett, 2011). This has been a source of controversy in the US as educators have argued that 'formative assessment is not a test but a process' (Popham, 2008, p6). This is compounded in the 'interim assessments'

found in American education. In these a period of instruction (eg five weeks) is followed by a formative test, the results of which are used for a short period of instruction which addresses errors on the test. The instruction in this period is designated 'formative assessment', though the evidence is that the test results are mainly used for reporting purposes (Perie et al., 2007).

The concept of *mastery learning*, with its emphasis on breaking tasks down into small steps and showing progressive mastery of them, is still prevalent in many areas of education. For example, the achievement of multiple discrete competences in vocational education is widespread (see Torrance, 2005; Sadler, 2005). The use of 'micro-teaching' to address detailed learning objectives in relation to examination preparation might also be treated as a form of this (Stobart, 2002).

Torrance and Pryor (1998) identified a number of formative assessment classroom practices in the UK which reflected behaviourist thinking. These 'convergent' formative assessment practices adopt a 'closed' approach to teaching and learning in which assessment 'starts from the aim to discover *if* the learner knows, understands or can do a predetermined thing' (p193). It is contrasted with 'divergent formative assessment' which seeks to discover *what* the learner knows, understands or can do. Convergent formative assessment is teacher controlled and is typified by closed questions and tasks, contrasting of errors with correct responses, and judgemental feedback focused on successful completion of the task.

5.2.5 Cognitive/constructivist[6] approaches to formative assessment

While the term *formative* may have come out of the behaviourist movement, other theoretical approaches have incorporated, and reinterpreted, some of the key concepts in ways that have extended the scope of formative assessment. Cognitive approaches rejected mechanistic stimulus-response accounts of learning and emphasised the role of individuals in constructing meaning for themselves. The focus is on 'how the mind works' – how information is processed and how learners 'make sense' of information. Shepard has summarised this as 'meaning makes learning easier, because the learner knows where to put things in her mental framework, and meaning makes knowledge useful because likely purposes and applications are already part of the understanding' (1992, p319). In this, the learner is 'more actively a subject as well as an object of formative assessment' (Brookhart, 2004 p2).

The theoretical basis of this approach, which in America loosened the grip of behaviourist thinking, is diverse. Eric Bredo (1997) identified some of the key elements as:

[6] These terms are interchangeable, *cognitive* reflecting common American usage and *constructivist* being more widely used in Europe.

1. A revival in the focus on 'mind' which had been central to William James' earlier pragmatist theorising (see below)
2. The development of artificial intelligence and studies of information processing (Simon, 1957)
3. Psychological/scientific studies of concept formation (Bruner, 1960).

Authoritative contemporary applications of this approach can be found in Bransford et al. (2000), *How People Learn: Brain, Mind, Experience and School;* Pellegrino et al. (2001), *Knowing What Students Know: The Science and Design of Educational Assessment*, and Pellegrino et al. (2012), *Education for Life and Work.* What is significant about these accounts is that they place little emphasis on the sociocultural aspects of learning.

In terms of formative assessment, the emphasis on the learner 'making sense' leads to an emphasis on clear learning goals, knowing the standard to be reached, and receiving feedback which closes the gap. The seminal contribution here is that of Royce Sadler in his highly influential 1989 paper *Formative Assessment and the Design of Instructional Systems.* This paper drew on cybernetic theorising from Norbert Weiner (1965, 1968) and Arkalgud Ramaprasad (1983) which emphasised how information could be incorporated into self-direction, with a system able to change its own direction (for more detail see Roos & Hamilton, 2008; Wiliam, 2011).

Sadler (2010) saw what goes on in the learning process, particularly feedback, as a far more complex cognitive process than in earlier theorising. The key to this was seeing the learner actively at the centre of the process, particularly in terms of understanding and experiencing the standard to be achieved. The teacher's role in this was to design the instructional system so that students could participate in the standard-setting process and make sense of the learning process. The emphasis here is on the development of self-regulated learning, 'progressive appreciation by the learner of what constitutes high quality work and the strategies needed to attain high standards, and hence high achievement' (p78).

This is reflected in much current practice and writing. It is embedded in Norway's Education Act, which articulated four key principles for effective formative assessment, holding that pupils learn best when they:

1. Understand what they are supposed to learn and what is expected of them;
2. Receive feedback that informs them about the quality of their work or performance;
3. Receive advice on how they can improve;

4. Are involved in their own learning activities for example through assessing their own work and development.

(Education Act Section 3-1, 3-11, 3-12, 3-13: Norwegian Directorate for Education and Training, 2011)

5.2.6 Socio-constructivist approaches to formative assessment

The limited role given to the social context of learning in cognitive approaches has led to a call for a more direct recognition of the social in the process of 'making sense' of information. The key statement of this came from Lorrie Shepard in her 2000 Presidential Address to the American Educational Research Association (AERA), which was published in 2000 as *The Role of Assessment in a Learning Culture.* In this she critiqued the legacy of behaviourist theory, particularly 'the continued intellectual kinship between one-skill-at-a-time test items and instructional processes aimed at mastery of constituent elements' (p5).

She saw socio-constructivism as 'borrowing from cognitive, constructivist and sociocultural theories' and acknowledged that 'these camps are sometime warring with each other' (Shepard, 2000, p5). The difference from sociocultural approaches (see below) is in foregrounding individual cognition while recognising the importance of culture and context. Her claim was that that 'this merged, middle ground theory will eventually be accepted as common wisdom and carried into practice' (p5). Other support for this synthesising of the constructivist and sociocultural comes from two classic papers, Paul Cobb's (1994) *Where is Mind? Constructivist and Sociocultural Perspectives on Mathematical Development* and Anna Sfard's (1998) *On Two Metaphors for Learning and the Dangers of Choosing Just One*.

Though Shepard did not make explicit connections, the socio-constructivist position could be seen as a descendant of the pragmatist/functionalist tradition of William James (1842–1910) and John Dewey (1859–1952). Functional psychology emphasised the transactional nature of learning. In this, learning is not simply shaped by the environment (as in behaviourism) or determined within the organism (constructivism) but is the result of continuous *transactions* between the environment and the individual which change both (Bredo, 1997). For education, this meant learning based on activity and reflection. Dewey argued that 'the teacher becomes a partner in the learning process, guiding students to independently discover meaning within the subject area' (Dewey, 1897), a view which overlaps with sociocultural theory.[7] These priorities align with those elements of formative assessment that emphasise active and social engagement, clarity about what is being learned and self-regulation.

There has been a distinct shift in recent European theorising on formative assessment towards more sociocultural understandings of formative assessment. *Sociocultural* is an

[7] A key difference between Dewey's approach and that of sociocultural theorists was his emphasis on the individual as a self-determining agent who helps to construct the environment – the basis of his views on democracy (Dewey, 1916).

umbrella term for approaches which root learning in social interaction. There are a variety of strands within this, the dominant one being based on the work of Lev Vygotsky (1978) and developments of this dialectical approach found in cultural historical activity theory (Engestrom, 1987). Others, for example Ecclestone (2002) and Pryor and Crossouard (2008), incorporated the sociological writings of such authors as Bourdieu (1990) and Bernstein (1996).

A largely separate strand in sociocultural thinking is that of the 'situated learning' of Lave and Wenger (1991), with its concept of 'community of practice' in which the learner is apprenticed into context-specific learning. While these concepts were derived from anthropology, they link to the American educational tradition of William James (1899, 1958), John Dewey (1916) and Jerome Bruner (1960), with their emphasis on applied learning and recognition of the social context of schooling (James, 2006).

The key commonality of these two strands is that of developing learner identities through social interaction. The emphasis is on the processes of negotiating roles and active social interaction, which in turn modify the teacher-learner contract, with assessment being done *with* the learner rather than *to* the learner. Motivation is not seen as separate from learning but as part of developing a learner's identity; 'being able to relate current learning both to future study and desired identities is potentially a powerful and useful way of looking at motivation through a social lens' (Pryor & Crossouard, 2008, p17).

These features are summarised in Table 1. Some of the positions of writers and practices are debatable given the overlaps between orientations.

5.3 Definitions of formative assessment and AfL

> But why is definition important in the first place? Definition is important because if we can't clearly define an innovation, we can't meaningfully document its effectiveness.
>
> (Bennett, 2011, p8)

The intention of this section is to present some working definitions of formative assessment which can be seen as representative of the different learning traditions previously described. Since the 1990s the term 'assessment for learning' has been preferred by many writers and practitioners, particularly outside the US, as it signals more directly the purpose and use of assessments intended to directly facilitate learning (see Wiliam, 2011 for a summary of the origins of the term). Though distinctions have

been made by some writers[8], the terms 'assessment for learning' and 'formative assessment' will be used interchangeably here.

Table 1 Summary of theories of, and implications for, formative assessment

Theoretical orientation	Associated with	Formative assessment emphasis	Typical practices
Behaviourist/ neo-behaviourist	Thorndike Gagne Bloom Popham (early) Test publishers	Atomised/step-by-step mastery Regular testing for error detection and correction Tests as formative assessments (product not process)	Learning objectives Tests to establish what is not known Test feedback to teacher to modify instructions Feedback to student corrective
Cognitive/ constructivist	Piaget Bruner (early) Simon Chomsky Bransford Pellegrino Ramaprasad Sadler Roos and Hamilton	Need for learners to 'make sense' of information and developmental schemas Importance of learner understanding Learning objectives and success criteria ('standard') Feedback as dynamic process (cybernetics)	Negotiated learning Intentions and success criteria Feedback information for learner to close gap Self-regulated and self-monitoring learning
Social constructivist	Crooks Shepard Cobb Sfard Assessment Reform Group Black and Wiliam (1998a)	Importance of school and classroom ethos Dialogue and negotiated learning Self- and peer assessment Motivation through engagement	Classroom expectations Encouraging learner engagement Active learning – dialogue, group work, self- and peer assessment
Sociocultural	Vygotsky Lave and Wenger Torrance and Pryor Pryor and Crossouard Black and Wiliam (2006) Ecclestone et al. (2012) Allal Perrenoud	Learner identity and changed teacher role and identity Negotiating understandings of task and quality criteria Apprenticeship model of learning Social context central to learning – classroom ethos (regulation)	Renegotiated learner identities Collaborative classrooms Learning through active social processes and interactions Changed classroom 'contract' around learning

[8] Black et al. offered the following differentiation: '*Assessment for learning* is any assessment for which the first priority in its design and practice is to serve the purpose of promoting pupils' learning…such assessment becomes 'formative assessment' when the evidence is actually used to adapt the teaching work to meet the learning needs (p2–3).

5.3.1 Behaviourist approaches

Bloom's initial definition of formative assessment, 'to provide feedback and correctives at each stage of the learning process' (1968, p48) was in the context of mastery learning, with a formative assessment being a test to identify what was known or not known. A formative assessment 'is designed to give students information, or feedback, on their learning…each student has a detailed prescription of what more needs to be done to master the concepts and skills from the unit' (Guskey 2010, p109). When this corrective work is done Bloom recommended that a second formative assessment be taken, using similar items to verify whether the 'correctives' worked and to offer a second chance – which also has motivational value. Linda Allal (2005) has described this approach as 'retroactive regulation'.

Within this tradition there has been a shift by some to see formative assessment as a *process* rather than a product. James Popham's (2008, p6) more recent definition of formative assessment is:

> A planned process in which assessment-elicited evidence of students' status is used by teachers to adjust their ongoing instructional procedures or by students to adjust their current learning tactics.

It is worth noting that assessment is still seen as planned (rather than informal moment-to-moment) and is firstly feedback to the teacher in order to modify instruction. The OECD (Lingard & Grek, 2005) definition shows some similarities: 'frequent, interactive assessments of student progress and understanding. Teachers are then able to adjust teaching approaches to better meet identified learning needs' (p13).

5.3.2 Constructivist and social constructivist approaches

The shift in definition for these learning theories is that the learner, rather than the teacher, moves increasingly into the foreground, with an emphasis on self-regulation. The focus is on the learner actively being part of the assessment process through understanding what is being learned, the quality of learning required and the ability to monitor performance against this. There is also more emphasis on informal, day-to-day assessment in contrast to planned assessment events. The Assessment Reform Group's (2002) definition, echoing Ramaprasad (1983), captured some of this:

> …the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there.

A more recent definition by Black and Wiliam (2009, p9) emphasised the contingent and inferential nature of formative assessment:

> Practice in a classroom is formative to the extent that the evidence about student achievement is elicited, interpreted, and used by teachers, learners and their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited.

The emphasis on formative assessment being informal and part of everyday classroom activity is captured by the joint definition of the third international conference on AfL in Dunedin, New Zealand:

> Assessment for Learning is part of everyday practice by students, teachers and peers that seeks, reflects upon and responds to information from dialogue, demonstration and observation in ways that enhance ongoing learning.
>
> (Klenowski, 2009, p264)

What is noteworthy about this definition is the emphasis on classroom processes and the lack of reference to more formal 'episodic' assessment.

5.3.3 Sociocultural approaches

There have been fewer attempts to define formative assessment directly in terms of sociocultural theory, though the need for this has been acknowledged (Allal, 2005; James, 2012). John Pryor and Barbara Crossouard in their *A sociocultural theorisation of formative assessment* (2008) interpreted formative assessment as 'a discursive social practice, involving dialectical, sometimes conflictual, processes' (p1). At the heart of this is the redefinition of roles within the social context of the classroom, in which both teachers and learners take on new *identities* in the learning process in which

> formative assessment…would make an explicit aim of raising students' critical awareness both of the discourses of the educational setting and also the wider social construction of these discourses.
>
> (Pryor & Crossouard, 2008, p10)

In practice this would involve negotiating understandings of the task and the quality criteria and collaboratively fashioning learner identities, their role in relation to tasks and criteria. In this the role of the teacher also changes as learning is co-constructed. Black and Wiliam's sociocultural account using activity theory (px) does not offer a related definition but works within their earlier definition.

5.4 Critique of formative assessment

Because formative assessment has been perceived as an educational 'good thing' many of the research claims made for it were initially largely unchallenged. Increasingly there have been more critical reviews of these claims, the most comprehensive and widely cited being Randy Bennett's (2011) *Formative Assessment: a critical review.* In this he identified six interrelated issues:

1. The definition of formative assessment/assessment for learning
2. The claims commonly made for its effectiveness
3. The limited attention given to domain concerns in its conceptualisation
4. The under-representation of measurement principles in that conceptualisation
5. The teacher support demands formative assessment entails
6. The impact of the larger education system.

This section uses this framework to examine some of the criticisms of the claims for formative assessment.

5.4.1 What does 'assessment for learning' mean?

While the issue of differing definitions has been discussed in an earlier section, Bennett highlighted a particular issue, the largely American debate about whether formative assessment is a product (eg a published test) or a process. One way around this confusion has been the switch to 'assessment for learning' by leading American writers such as Rick Stiggins (2005). However, Bennett suggested that the switch to assessment *for* learning and assessment *of* learning (Gipps, 1994)[9] brings its own problem as it 'absolves summative assessment from any responsibility for supporting learning' (p7) – despite the capacity for summative assessment to play an important role in learning. For example, preparing for good quality tests can be a positive learning experience. Bennett's own position is:

> Formative assessment then might be best conceived as neither a test nor a process, but some thoughtful integration of process *and* purposefully designed methodology or instrumentation. Also calling formative assessment by another name may only exacerbate, rather than resolve, a definitional issue. (Bennett, 2011, p7)

[9] Others had used the term 'assessment for learning' earlier (Mittler, 1973; Black, 1986; James, 1992), but it was Gipps who introduced the contrast between *for* and *of*.

5.4.2 Formative and summative issues

The complex relationship between formative and summative assessment has been an ongoing concern. The general consensus (eg Bennett, 2011; James & Harlen, 1997; Harlen, 2007; Black et al., 2003) is that the two are interrelated and that assessment developed to serve a summative function can be used formatively – though some formative assessment may not be appropriate for summative use. In cultures which are examination dominated, such as Hong Kong, assessment serving a formative function will need to be aligned to the focus of high-stakes settings before teachers are likely to adopt it (Johnson et al., 2008; Carless, 2011).

Maddalena Taras took a different position in her 2009 article, 'Summative assessment: the missing link for formative assessment', where she argued that summative assessment is at the heart of formative feedback since the teacher has to make a summative judgement on where the learner is before any formative processes can take place. Her argument is that it is a mistake to discuss assessment in terms of *functions* (formative, etc.) rather than *processes*, since formative assessment involves processes, which in terms of functions are both formative and summative.

A direct break from this consensus is that by Roos and Hamilton (2005), who argue that it is 'unhelpful to treat them as opposite sides of the same thing' (p18). Using a cybernetic model they see them as different paradigms, reflective of the original distinction made by Scriven.

5.4.3 Exaggerated claims for effectiveness

There is a body of literature claiming that AfL, or formative assessment, is one of the most powerful tools to increase students' learning (Bangert-Drowns et al., 1991; Black & Wiliam, 1998a, 1998b; Brookhart, 2004, 2007; Crooks, 1988; Dempster, 1991, 1992; Elshout-Mohr, 1994; Fuchs & Fuchs, 1986; Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Köller, 2005; Natriello, 1987; Shute, 2008; Wiliam, 2007). The main ideas from the literature focus upon how students can use different forms of feedback and assessment to improve their learning, from peers, teachers or others, as long as it is of high quality and timely. Self-assessment is also included in the AfL processes, a procedure that demands an active student role, but also helps students to gain more insight into their learning and make them more responsible for their own learning outcomes (Stobart, 2008). There seems to be consensus in research regarding the positive effects of AfL, as perceived by research participants (see, for example, Ofsted, 2008; DfES, 2007; Condie et al., 2005; Hayward & Spencer, 2010; Kirton et al., 2007; Webb & Jones, 2009; Kellard et al., 2008, among others).

Most of the claims about the effectiveness of formative assessment are based on estimates reported by Black and Wiliam in their 'Inside the Black Box' pamphlet

(1998b). In 'Inside the Black Box', Black and Wiliam (1998a) estimated that feedback would increase students' learning within the range of effect sizes from 0.4 to 0.7, while the review by Shute (2008) found effect sizes between 0.4 and 0.8. In the review of 74 meta-analyses, conducted by Hattie and Timperley (2007), an average effect size of 0.95 standard deviations was reported from an analysis of 4,157 studies. Other writers reported learning gains with effective sizes between 0.4 and 0.7 (Popham, 2008b; Stiggins, 1999). Given that Black and Wiliam's (1998b) review was not a meta-analysis, [10] but is generally reported as one, Bennett described this as a 'mischaracterisation that has essentially become the educational equivalent of urban legend' (p12). He also critically reviewed widely quoted reviews by Nyquist[11] (2003), Meisels et al. (2003) and Rodriguez (2004). A similar critique of the OECD's *Formative Assessment* (Looney, 2005) in the 'What Works' series is made by Jannette Elwood (2006).

The meta-analysis critique was further developed by Neal Kingston and Brooke Nash (2011). They argued that Black and Wiliam's estimate has been treated as a meta-analysis, which it is not, and conducted their own meta-analysis. They used five inclusion criteria for studies: the description of the intervention as *formative* or *assessment for learning*; an academic K–12 setting for participants; use of a control or comparison group design; inclusion of appropriate statistics to calculate effect size; a publication date of 1988 or after. This yielded only 13 studies, which among them involved 42 effect sizes (with two studies contributing 27 of these – including using 21 individual teacher effect sizes from the Wiliam et al. (2004) study). Across these studies they found a weighted mean effect size of 0.28 – well below the often-cited 0.4–0.7 from Black and Wiliam's estimates. They acknowledged that even this effect size has 'great practical significance in today's accountability climate' and called for clearer descriptions 'of the form and features of formative assessment' (p35). Specifically, they called for the provision of effect sizes in evaluation studies, for more representative sampling and for more focus upon the specific factors influencing the efficacy of formative assessment.

Kingston and Nash's review was in turn critiqued for its methodology by Derek Briggs and colleagues (2012). They had concerns about the narrow initial search for studies and about inconsistencies in what was selected. Some of the variability in effect sizes could be explained by the different calculation techniques, it was argued. They concluded that as a consequence, even the 0.20 estimate cannot be claimed with confidence – the effect sizes from a different sample might have been significantly different from than this.

[10] Interestingly John Hattie's (2009) much larger meta-analysis does come up with high effect sizes for teacher feedback (0.72) and metacognition strategies (0.67). Hattie´s use of effect sizes have been criticised by several scholars, including in Norway. See Imsen (2011) for a review in Norwegian.

[11] Though this was only because it was not published in a peer-reviewed journal.

5.4.4 Research design

A dilemma is that much of the research on formative assessment has not used designs which allow for pre- and post-test measures which would allow for direct claims about effectiveness (Stobart, 2010). This is compounded by the range of elements included in current understandings of formative assessment which make it hard to pin down what has caused any changes. This criticism was been developed by Karee Dunn and Sean Mulvenon (2009), who argue that the 'vagueness of the constitutive and operational definitions directly contributed to the weaknesses found in the related research and dearth of empirical evidence identifying best practices related to formative assessment. Without a clear understanding of what is being studied, empirical evidence will more than likely remain in short supply' (p2).

Bennett (2011, p8) argued that what is needed here is a theory of action which

> identifies the characteristics and components of the entity we are claiming is 'formative assessment', along with the rationale for each of these characteristics and components; and…postulates how these characteristics and components work together to create some desired set of outcomes.

This would then need to be followed by applications that illustrate what this would look like in practice. In terms of validity two types of argument would be needed (Bennett, 2011, p14):

> a *Validity Argument* to support the quality of inferences and instructional adjustments, and an *Efficacy Argument* to support the resulting impact on learning and instruction.

5.4.5 General versus domain-specific strategies

Earlier formulations such as Bloom's treated formative assessment as an instructional approach, independent of the subject being taught. This was echoed in some of the later formulations of formative assessment, for example Black and Wiliam's 'Inside the Black Box' (1998b) and the Assessment Reform Group's *Assessment for Learning: Beyond the Black Box* (1999). We speculate that this aligns with constructivist views of general learning strategies. It has increasingly been recognised that the subject context will modify formative practices. An example of this was the move to subject-specific 'Inside the Black Box' pamphlets.[12]

[12] Their initial action research, and subsequent publications, focused on secondary maths and science teaching, with English incorporated later. However their 'Inside the Black Box' was extensively used as a general guide to AfL practices.

Bennett (2011, p15) claimed that to be maximally effective, formative assessment requires the interaction of general principles, strategies, and techniques *with* reasonably deep cognitive-domain strategies.

This is reflected in Black and Wiliam's move to a more explicit sociocultural position (2009) in which they emphasised the importance of knowledge of pedagogical content (Shulman, 1986).

Coffey et al. (2011) in their *The Missing Disciplinary Substance of Formative Assessment* make this point in relation to formative assessment in science. Their concerns are that 'in its concentration on *strategies for the teacher*, the literature overlooks the *disciplinary substance* of what teachers and students assess' (p1). They called for a shift in which 'formative assessment moves out of strategies and into classroom interaction, with roots in disciplinary activity and goals' (p23). Anne Watson (2006) made a similar point in relation to maths teaching.

5.4.6 The measurement issue

Bennett's (2011, p17) claim was that the formative assessment literature gave too little attention to the *interpretation* of assessment evidence. The process of finding out what a learner knows and understands involves a formative hypothesis that requires further validation through the collection of further evidence. Such inferences are uncertain and are

> subject to systematic, irrelevant influences that may be associated with gender, race, ethnicity, disability, English language proficiency, or other student characteristics. Put simply, a teacher's formative assessment may be unintentionally biased.

Elwood (2006) offered a more radical sociocultural critique of possible biases in interpretation. She observed that little attention has been paid to equity and how the change in the teacher–student relationship plays out in terms of fair assessment. Using the example of gender she argued that the focus on the individual leads to a neglect of 'the gendered nature of students' lives and experiences' (p228). This leads to a broader critique of constructivist positions in which considerations of the social and cultural aspects and settings of classrooms seem more rhetorical than actual and she argued that while lip-service might be paid to the notion that formative assessment can deliver equality of outcomes, the work reported was not substantial enough to support these claims (p228). This leads to the argument that there had been no real 'paradigm shift' in relation to assessment; formative assessment remained 'a consideration of what the student can do alone after learning through social interaction' (p230), leaving us with a symbolic view of cognition and a mind still located 'in the head' (Cobb, 1999, p135).

Elwood's (2006, p230) own sociocultural formulation is that teaching and learning are essentially about *interactions*:

> …learning as a cultural activity, mind as between individuals and assessment and testing as cultural activities that can only describe students' learning in relationship to the assessment tasks, their teachers and their experiences.

5.4.7 The professional development issue

Formative assessment requires teachers to have pedagogical knowledge of how to teach effectively. Bennett (2011) added to this two other neglected features: 'reasonably deep cognitive-domain understanding and knowledge of measurement fundamentals' (p18). He argued that these two features were seen as optional extras in formative assessment professional development programmes, rather than essential elements.

5.4.8 The impact of the larger educational system

There is increasing recognition of the impact of the broader educational and social systems within which formative assessment functions. How well formative assessments align with the other elements of the system is critical. A key concern within this is the importance attached to the accountability system and the role of summative assessments.

Kennedy et al. (2008) developed this in relation to Hong Kong. They argued that in a Confucian heritage society examinations are likely to be central to teaching and learning and that proponents of formative assessment must recognise that, to gain any credibility with teachers and the broader society, it needs to be related to summative assessments. John Biggs (1998), himself deeply familiar with Confucian heritage society, made the same point in his response to Black and Wiliam's 1998a review:

> My main substantive problem is with the decision to exclude SA [summative assessment] from a review of the effects of assessment on classroom learning…These effects, referred to as 'backwash'…are usually seen as entirely negative, and interestingly, as stronger than the positive effects of FA [formative assessment] ('feedback'). This suggests clearly that significant gains are to be found as much in mitigating or reversing backwash as by enhancing feedback.

Kennedy et al. (2008) adopted this approach, reasoning that 'the consequential validity of assessment practices can be determined by the extent to which a specific practice, whether it is formative or summative, provides the conditions for the students to learn'

(p204). It therefore becomes important to improve the quality of summative assessment as well as adapting formative assessment, which is 'very much a Western construct' (p198), to local needs such as quality feedback from classroom summative assessments.

This is one example of how cultural and educational systems influence the implementation of formative assessment. Bennett concluded that 'we have to rethink assessment from the ground up as a coherent system, in which formative assessment is a critical part, but not the only critical part' (2011, p20).

5.5 Implementation of AfL

AfL can be considered nowadays as a 'research epidemic' that has in the last decades 'feverishly spread into every discipline and professional field' (Steiner-Khamsi, 2004, p2). It has been researched, piloted or implemented in a wide variety of contexts (Australia, New Zealand, the US, Canada, Hong Kong, Chile, Singapore, Rwanda, Cameroon, Malaysia, and the Netherlands, just to mention a few), usually in the context of professional development programmes that involve close cooperation between teachers and researchers (see for example Tierney & Charland, 2007; Torrance & Pryor, 2001; MacPhail & Halbert, 2010; Condie et al., 2005; Dori, 2003; Hayward & Spencer, 2010; Azúa & Bick, 2009, among many others) as well as guidelines for the implementation of the approach as a policy.

A review of the literature on the implementation is given in Hopfenbeck et al. (2013), and highlighted the following issues. Many teachers are reluctant to use peer and self-assessment because they do not fit the dominant model of teaching and learning, especially the expectations about the roles of the teacher and student (MacPhail & Halbert, 2010; Willis, 2008; Kirton et al., 2007; Stiggins & Arter, 2002). So it was argued that researchers underestimate the challenge of transforming from monologic teaching to a more dialogic teaching approach Dysthe (2008). Such change in teachers' and students' roles is essential for AfL, and without these changes, students will not use feedback to learn (Harlen, 2007, p20). These processes are challenging, but teachers can foster self-regulated learners when using AfL practices (Hopfenbeck, 2011). It is acknowledged, though, that it takes time and sustained effort to support teachers to foster such learning environments (Perry et al., 2007).

As mentioned in section 5.4.5, Black and Wiliam (2009) emphasised the importance of pedagogical content knowledge (Shulman, 1986) when implementing AfL. Concerns about lack of subject content knowledge and assessment skills among teachers have been raised by a number of authors (eg DfES, 2007; Kellaghan, 2004; Carless, 2005; Gioka, 2006; Stiggins & Arter, 2002; Ofsted, 2008; Thompson & Wiliam, 2008; Azúa & Bick, 2009; Gardner et al., 2011). Another concern related to whether teachers were part of a whole-school commitment to the implementation. Black et al. (2010) referred to the

importance of creating a culture of assessment and of generating a change at a whole-school level. In the UK, Ofsted reports have observed detachment from senior staff in relation to AfL in practice, although research considers this a crucial factor if AfL is to be implemented in the long term (Ofsted, 2008).

A further challenge is the confusion around the theory underpinning AfL, which can lead to a superficial understanding of the approach of AfL. This might lead to a mechanical or equivocal use of AfL in practice (Hayward & Spencer, 2010; Tapan, 2001; Webb & Jones, 2009; Azúa & Bick, 2009; Harlen, 2007). Black and Wiliam (1998a) also argued that teachers would not take up attractive-sounding ideas, albeit based on extensive research, if these principles were presented as general principles that leave the task of translating them into everyday practice entirely to the teachers. The everyday classroom is too busy and Black and Wiliam suggested that transforming general principles into AfL practices would only be possible for the outstanding few. It is also acknowledged that replicating small-scale studies, which are often conducted by dedicated teachers, is more challenging on a large scale. The questions raised here relate to the extent to which the complexity of support and the ideal conditions of small pilot programmes can be replicated when the approach is to be disseminated at a wider level (Torrance & Pryor, 2001; Dori, 2003; Black et al., 2004; Black et al., 2010). Another challenge is that high-stakes testing systems sometimes conflict with priorities in AfL, and there is a mismatch between the AfL approach and the high-stakes testing and accountability systems (Black et al., 2010, 2004; Condie et al., 2005; Dori, 2003; Hayward & Spencer, 2010; Kirton et al., 2007; Gipps et al., 2005; Kellaghan, 2004). Teachers' values can be in conflict with such systems, and a feeling of deprofessionalisation can occur (Shepard, 1992; Berryhill et al., 2009). Teachers' values are also linked to their beliefs about assessment and learning, and several authors have written about the importance of addressing teachers' and students' beliefs on assessment and learning as a relevant part of the change management implementation process of formative assessment strategies (see, for example, Torrance & Pryor, 2001; Brown et al., 2009; Brookhart, 2001; Willis, 2008; Tapan, 2001; Carless, 2005; Marshall & Drummond, 2006).

Despite the empirical evidence for positive effects of AfL upon learning, the literature suggests that *how* formative assessment is designed, developed, embedded and implemented by teachers is poorly understood (Ayala et al., 2008; Pellegrino et al., 2001). Some of the challenges discussed in the literature relate to the lack of a coherent theoretical understanding of AfL and formative assessment, lack of empirical studies supporting the strong claims and lack of an agreed-upon lexicon with regard to the term 'formative assessment' (Dunn & Mulvenon, 2009; Bennett, 2011; Elwood, 2006). We turn next to look at how the well-established field of psychometrics has characterised learning and how it has contributed to our understanding of student learning.

Chapter 6 Views of learning in the psychometrics tradition

Jo-Anne Baird
OUCEA

In one sense, the psychometrics tradition has nothing to say about learning. Learning is not the concern of this field. Psychometrics as such, in so far as it relates to educational, psychological and social measurement in general, is concerned only with the status of the individual in terms of a trait or construct, and not directly with how the person achieved that status. By analogy a survey of measurement of the skinfold composition of body fat may be conducted because of a concern about obesity, but the measurement of the body fat density in itself does not explain how a person arrived at that skinfold density level, or how the person might reduce or increase that density. In the cognitive domain this means that psychometrics is concerned with knowledge and understanding, and not how that knowledge or understanding has been learned. Thus it does not assess the process of learning any more than a skinfold body mass index assesses the diet and exercise of a person that have led to that skinfold measure. Of course, in many cases some inference from a skinfold measure can be made as to what the diet might be, but that inference comes from knowledge of physiology; similarly in many cases some inference from a cognitive domain measure can be made as to what kind of learning has or has not taken place, but that inference comes from knowledge of learning theories. Therefore, the degree to which this inference can be made will be a function of the quality of the assessment to provide that inference. To create such assessments, deep study of the substantive field is required and goes beyond that which psychometrics as such can offer.

Psychometric models are statistical techniques that can be applied to content that is broader than educational assessment and may be unconnected with learning. Assessment of personality, for example, can be underpinned by psychometric techniques, but higher scores do not indicate learning. Psychometrics can be used with many variables, or constructs, with higher scores simply indicating a greater quantity of the construct. These constructs can be categorised as educational (eg examination scores), psychological (eg depression) and attitudinal (eg political). As Markus and Borsboom (2013, p45) pointed out, the meaning of the numbers generated by psychometric models is conjured by people generating the links with the substantive topic. The very act of measuring the status of a substantive construct has an impact upon how that construct is perceived. The story of the relationship between psychometrics and learning does not end here because the use of psychometrics with educational assessment data entails certain requirements and assumptions that have

implications which can frame the context and definition of learning outcomes – albeit implicitly or explicitly. Further, we will also discuss the extent to which psychometrics has contributed to, or could contribute to, our understanding of learning. There are three broad sets of psychometric models, which are described briefly in turn. We follow Borsboom's (2005) classification.

6.1 The psychometrics tradition

6.1.1 Representational measurement models

These models were formulated to address the questions of whether psychological phenomena are quantifiable and what is meant by measurement in psychology as opposed to, for example, physics. Stevens (1946) broadened the notion of what counted as measurement by defining it as 'the assignment of numerals to objects or events according to rules' (p677). The idea is that our measurement scales represent the attributes of interest through mathematical structures and relations. This work produced the widely influential classification of measurement scales as nominal, ordinal, interval and ratio, indicated what the mathematical relations between the attributes and scales are and noted which statistical procedures could be used with each scale (Stevens, 1946; Table 1, p678). What kind of scale we have is a function of the attributes *and* the measurement instrument (p677-8). In one fell swoop, Stevens defined measurement as a product of the testing instrument and procedures. The formal models, or equations, depend upon the scale that can be constructed, which Stevens argued was limited by the properties of the attributes of interest and may, for example, only be on a nominal scale. Michell (1997) contested this approach, arguing that it was an unscientific view of measurement that side-stepped the prior problem of whether psychological attributes are quantifiable at all.

In their famous paper on additive conjoint measurement, Luce and Tukey (1964) showed that once certain conditions were satisfied, ratio scales could be produced from transformations of the combination of different ordinal variables (known as additive conjoint measurement). This exciting development has not been taken up to a large degree in psychology or education (Cliff, 1992) because we can see empirically that the conditions are frequently not met. For example, test parameters change across people and across time, so the requirement for transitivity is not met (Markus & Borsboom, 2013, p36). The foundations of this approach are philosophically and mathematically sound, but its relationship with data is problematical. In the less strict sense in which Stevens (1946) formulated representational measurement, it is the dominant measurement theory in use in psychology and education. Some latent trait models (one-parameter or Rasch models) are consistent with additive conjoint measurement (Andrich, 2003, p559–560).

6.1.2 The classical test theory model

Classical test theory comes from centuries of research on error (Porter, 1986, p7) and is a formalisation of the observation that all measurement entails error. In Equation 1, X represents the observed score, which is composed of a true score (T) plus error (E). The mean of repeated measurements tends towards the true score, as error has a mean of zero and is normally distributed unless there is bias. By definition, the true score is the average of repeated measurements. Classical test theory took a leap forward in its theorisation and formalisation in the publication by Lord and Novick (1968). Generalisability theory (Brennan, 2001) is an extension of this approach that decomposes causes of variance further.

$$X = T + E \qquad \text{(Equation 1)}$$

6.1.3 Latent trait models

Latent trait models are related conceptually to factor analytic models first introduced by Spearman (1904). Algebraically, the simplest is the Rasch (1960) one-parameter model. In Equation 2, the probability that a person will get an item correct is a logistic function of the difference between a person's ability (β_n) and the difficulty of the item (δ_i). In this particular model, a person's ability and an item's difficulty are conceived as scores on the same trait. A person is predicted to give the correct response to an item if their ability exceeds its difficulty and vice versa. Two-parameter models are also available. The extra parameter takes into account the level of discrimination or the extent to which the item separates the able from the less able. Three-parameter models are more widely used than two-parameter models; the third parameter deals with the possibility of guessing on multiple-choice items.

$$\Pr\{\text{positive response}\} = \frac{e^{(\beta n - \delta_i)}}{1 + e^{(\beta n - \delta_i)}} \qquad \text{(Equation 2)}$$

A practical consequence of some latent trait models is the possibility of an incomplete design, in which every student takes every item. For example, in international tests such as PISA, students take only a subset of the items and their ability is measured on the basis of this. So long as there is a way to link people parameters (through common items) and/or item parameters (through the same people taking particular items), an incomplete design is sufficient to estimate all of the ability and item difficulty parameters. Therefore efficiency is a benefit of latent trait models.

6.1.4 The contribution of psychometrics to learning theory

No general learning theory was born from a psychometric approach. Behaviourist, cognitive, socio-constructivist and sociocultural theories did not depend upon psychometrics. Nor has psychometrics been a major research technique in the field of

learning theory research and development.[13] This disconnect is striking, as many authors espouse the considerable benefits and advances made by psychometrics and deride the lack of uptake of these sophisticated techniques in psychology and education (eg Cliff, 1992; Andrich, 2004; Borsboom, 2006). Some argue that lack of statistical skill explains, at least in part, why psychometrics has not been used more extensively (Cliff, 1992; Pellegrino et al., 2001, p168; Borsboom, 2006). Pellegrino et al. (2001, p111) wrote:

> The measurement models in use today include some very sophisticated options, but they have had surprisingly little impact on the every day practice of educational assessment. The problem lies not so much with the range of measurement models available, but with the outdated conceptions of learning and observation that underlie most widely used assessments.

Perhaps if quantitative skills were in abundance amongst educationalists, psychometrics would have had a larger role in the history of learning theory, but there are other major causes for the disjuncture between the fields of psychometrics and learning theory.

6.1.5 Empirical contribution of psychometrics to understanding learning

Interaction between data and theory in an inductive-predictive paradigm is not the only way in which advances in theory germinate (Wolpert, 1992). Nonetheless, testing theory through falsifiable empirical methods is a standard scientific methodology (Popper, 1959). Wilson (2005) has made significant attempts to use psychometric techniques in this way, using his Berkeley Evaluation and Assessment Research system. This approach entails documenting the construct to be learned and what constitutes progression in this construct (Wilson, 2009). Items are written to test the construct at different levels and analysis is conducted using psychometrics to see whether the items have operated as intended. More difficult items are evidence of higher levels of progression. Of course, the items will not fall in a standard pattern of difficulty for every student who takes the test. This is a statistically-based, normative model which describes the general pattern. Where items do not conform to the predetermined order of difficulty, there may be something wrong with the original ideas of progression in the construct or there may be something wrong with the task design (Wilson, 2006, p654). Andrich (2004) argued that these anomalies should be of great interest because they challenge the current theory of learning relating to the construct.

These anomalies occur fairly frequently. For example, Wilson (2006, p654, Figure 4) reported that an item relating to recognition in Task 6 turned out to be easier than expected. Our models of construct progression are far from perfect and we are not good

[13] Although there are examples of the use of psychometrics as a tool to investigate learning theory, mainly from a cognitive perspective (eg Embretson, 1993; Mislevy, 2009), this has not been the main methodological paradigm.

at predicting how difficult items will be for students (Brandon, 2004; Good & Cresswell, 1988; Impara & Plake, 1998; Wolf et al., 1995). It is hard to reject the idea that we can use data to get better at this, even if research has shown that giving feedback to people trying to make judgements of the difficulty of items has not been entirely productive (for a review see Reckase, 2001). A fundamental problem is that the assessment instrument intervenes between the individual's knowledge and our view of that knowledge. Easy and difficult questions can be written about *the same* concept. For example, Briggs et al. (2006) provided a construct map for understanding the earth in the solar system. At level 4 (expected level for fifth-grade children), the student should know, for example, that the earth orbits the sun once a year. Figure 5 shows two example items that could be written on this topic, with the first having been made deliberately easy and the second difficult for the purposes of illustration.

Figure 5 Items on the earth's orbit of the sun

Easy item:

How long does it take for the earth to orbit the sun?

a) 1 day
b) 28 days
c) 365¼ days
d) all of the above

Difficult item:

What was Copernicus' theory about the motion of the earth and the sun? How long does the orbit take? How do we know that Copernicus, and not Ptolemy, is correct?

Items can be made more difficult in a number of ways: through the language used, by requiring more in-depth knowledge, making higher-order thinking skills a requirement, changing the scoring mechanism and so on. Outside the remit of the test designer, whether the student has been exposed to the test format has an influence upon the apparent difficulty of the item (eg Brunner et al., 2007). Curriculum exposure will also have a large effect; whether children were taught about Copernicus and Ptolemy would have a huge effect upon the difficulty of the second item. Our choice of item relates heavily to what we believe the curriculum aims are. If the curriculum aim is simply for children to know the fact that the earth rotates around the sun in 365¼ days, we would be content with the first item. Alternatively, if the curriculum aim were to develop scientific thinking more generally, then we would consider it very important to introduce the ideas that different theories can be put forward for the same phenomenon and that data can be used to choose between them. Both of the items in Figure 5 could

be valid, but they could not be used interchangeably to indicate progress on the construct of interest. In their book *Developing and Validating Test Items,* Haladyna and Rodriguez (2013, p28) indicated that the cognitive demand that is assigned to a test item is 'only a best guess based on the speculation about the cognitive process needed to respond to the test item correctly'.

One strategy has been to use 'standardised tests', which have been pre-tested, so that the item difficulties for the target population of test-takers are calculated in advance. Items can be held securely, so that students cannot easily learn how to perform on the specific items and thereby make them easier. Use of standardised, secure tests with multiple-choice items has been prevalent in the US.[14] These strictures make for a more standard measuring instrument, but it has been argued that they undermine the learning experience (Goldstein, 1979, 1980). Assessment dominates the learning experience in many countries because of the powerful effects of high-stakes testing (Madaus et al., 2009). It follows from these points that we need to better understand the impact of assessment design upon students' performances and several researchers are pursuing this topic (eg see Behrens et al., 2012; Ahmed & Pollitt, 2007).

Our conclusion has to be that although there are limitations, psychometrics has contributed empirically to understanding of which kinds of items and concepts tend to be more difficult for students than others. However, psychometricians rarely work closely with subject matter experts, so the data have not always been used as feedback to those who develop theories of learning. Further, where test items prove to be easier or more difficult than was anticipated, there is an over-determination problem, with one item representing many possible (design feature) causes for the unexpected pattern of results.

Psychometrics has traditionally been associated with multiple-choice testing formats. The standard psychometric techniques suit this format best. However, this limitation is really a thing of the past, theoretically if not pragmatically. Many techniques are now available to handle a range of assessment formats (see Pellegrino et al., 2001, chapter 4 for a review). The expertise in using these techniques may be more limited, which is why many have bemoaned the lack of application of psychometrics (eg Borsboom, 2006). Pellegrino et al. (2001, p169) concluded that despite the considerable advances in psychometric models, the most serious barrier to representing learning through these techniques was the lack of close collaboration between scientists, educators, task designers and psychometricians.

6.1.6 Defining the construct – whose job is it?

There has been angst over the level of contribution to knowledge about learning from the field of psychometrics generally (see Blinkhorn, 1997), but some researchers see

[14] Additional reasons for this include the fact that there are different curricula in each state and some authors cite commercial reasons for this because this test format is more economically efficient (Goldstein, 2012; Mislevy, 1997).

this as the job of substantive theory (Borsboom, 2006) and others see it as a collaborative enterprise (Andrich, 2004). Models of learning that underpin test design are often either not referred to or remain in 'puberty, infancy or even at the foetal stage' (Sijtsma, 2006, p453). Given the powerful use of assessments to define what counts as learning for many important societal functions, this is striking.

There are, of course, different ways of conceptualising theories of learning. Earlier, we discussed the broad theories of learning. Then we looked at how Mark Wilson is attempting to create depictions of progress in a construct that is a theory of learning within the cognitive tradition. Other examples of this occur where empirically-derived performance descriptors have been produced for standards-referenced examinations (Cizek et al., 2004). These constructs are often driven by a blend of theory and empirical findings about the test scores. Thus, the written descriptors form a working model of what it means to learn – a theory in practice.

As the connection between theory, data and the construct can be somewhat loose in practice, the status of the written constructs, as models of learning, is questionable in these instances. Lauder et al. (2006) described governmental attempts to produce constructs in this manner as a 'state theory of learning'. This approach typically has little to do with psychometrics, however. The intersection between substantive theories, assessment design and psychometrics seems to be a promising place in which to try to advance our understanding of learning.

Yet psychometrics is not a neutral tool in our quest to understand learning. Psychometrics is geared more towards outcomes than processes, so our theories of learning would be shaped in that manner. As noted earlier, our very definition of what it means to have learned would be moulded by assumptions underpinning the psychometrics tradition and its associated technologies. Psychometrics brings with it a set of assumptions about test formats, processes and procedures that educationalists might find unhelpful. This was the basis of the 'assessment' movement outlined earlier (Gipps, 1994). As discussed previously, Baird and Black (2013) argued that educational purposes need to be forefronted to a larger extent if test theory is to move forward. The compromises that are implied for test design by 'technocratic' psychometrics approaches (Lawn, 2008) currently take assessment and learning in the wrong direction for some. Gergen and Dixon-Román (2013, p7) wrote that

> …traditional measurement practices have been useful for certain groups in terms of providing a vantage point for deliberating about educational standards and policies. More debatable is whether tests have been successful in rending the educational system effective in attaining its goals.

Whether or not the reader is content with the direction in which psychometric principles take education, it is important to note their impact. Moss et al. (2005, p70) wrote:

> Different methods and theories have implications for the ways in which concepts such as learning or educational reform or fairness are formulated, studied and promoted through practical activity. Perhaps more profoundly and subtly, these methods and theories affect the ways human beings are represented and, ultimately the ways they come to understand themselves and others…

6.1.7 Dilemmas and the psychometrics tradition

6.1.7.1 Functionalism versus traits

The psychometrics tradition is divided on these issues. Operationalists, even in the latent trait approach, argue that the construct is whatever the trait measures. That is, the construct is made up of the items and no more. Markus and Borsboom (2013, p37) argued that this is not a sustainable approach because every test would be measuring something different and it would make no sense to try to equate results on different tests, or talk in general terms about scores in science tests, for example. They argued that the only way psychometrics can be understood is to see tests as measuring a latent trait. Notwithstanding their arguments, the latent traits are not always well conceptualised in educational assessment. New latent traits are generated frequently when new syllabuses and examinations are introduced. However, rarely do psychometricians stake a claim about whether the attributes of interest in the latent trait are thought to exist independently of the measure itself. Even where psychometricians believe that the latent trait exists, they may believe its existence is in the form of a social construct, in which case reality is of a different, postmodern form. Then again, a psychometrician could take the perspective that he or she is an engineer, not a scientist (Rust & Golombok, 1989, p26) and be agnostic about the physical or social existence of the attributes of interest.

6.1.7.2 Cognitive versus sociocultural

As we have seen, there has been an emphasis upon cognitive theory in psychometrics, although some would argue that it is still essentially behaviourist in its approach. Indeed, it is difficult to see how psychometrics can embrace sociocultural approaches to learning. As Moss et al. (2005, p77–78) argued, the standardised methods germane to psychometrics are ill-suited to elucidating social effects. Rather, they have been seen as a part of neoliberal social engineering (Moss et al., 2005, p77; Gergen & Dixon-Román, 2013). Exactly how psychometrics, or assessment more broadly, can tackle the social is unclear precisely because standardised approaches are part of the logic of fairness.

6.1.7.3 Positivism versus interpretivism

The collaboration between substantive theorists and psychometricians called for above could potentially operate in an interpretivist mode, investigating how children responded to different forms of assessment and producing theories of learning in this manner. However, psychometrics is fundamentally about the numbers assigned to the qualitative tasks and responses of people taking those tasks. It is a positivist methodology. Borsboom (2005) argued that classical test theory and latent trait theories were fundamentally realist, assuming that the traits we measure are pre-existing properties of individuals and not social constructions. Taking this to its logical conclusion, learning means getting a higher score on our tests. This places a heavy burden upon the construct being measured, as it is, in effect, a theory of learning. Note that this is the route that Mark Wilson has explicitly been elaborating in his work.

6.1.7.4 Realism versus postmodernism

Borsboom (2005) argued that the classical test and latent trait approaches were necessarily underpinned by a realist approach. However, since they rarely tell us in writing, without asking psychometricians, it is hard to tell what their philosophical positions are. Postmodern psychometricians might use the techniques whilst understanding that the latent traits may or may not exist. If you believe that the constructs are socially created, then it is possible to use the techniques of psychometrics with a view to producing a score that is related to other test scores *via* a latent trait. But you may also sustain the belief that the latent trait is socially constructed, not real (in the sense of existing independently of the measure), and that all you are *really* doing is generating items that correlate with each other, which has its uses.

We now turn to the literature on international tests, which have been developed based upon psychometric techniques.

Chapter 7 International tests

Therese N. Hopfenbeck and Jo-Anne Baird
OUCEA

Although international tests were invented in the last century, it is in the 21st century that they have really taken off in terms of their impact upon research and policy. The OECD, International Association for the Evaluation of Educational Achievement (IEA), United Nations Educational, Scientific and Cultural Organization (UNESCO), the World Bank and other intergovernmental organisations have a powerful effect upon policy-making across nations. The OECD, for example, has a wealth of technical expertise in test construction and analyses of the findings, as well as the know-how to link this with policy and, to some extent, to tailor their findings to particular contexts. In instances where OECD staff themselves lack expertise, they have well-established processes in which they create powerful networks of researchers and policy-makers. Whilst education is primarily seen as an area in which national policies prevail, some analysts have suggested that the technical authority of the OECD and their policy activist stance could create a global education policy field (Lingard & Grek, 2005). Thus, educational policies within a country need to take into account the processes of globalisation.

Today, Norway participates in the following studies initiated by the IEA: Progress in International Reading and Literacy Study (PIRLS), International Computer and Information Literacy Study, International Civic and Citizenship Education Study 2009 and TIMSS. In addition, Norway participates in the following surveys initiated by the OECD: Assessment of Higher Education Learning Outcomes (AHELO) and PISA. The impact of these studies has been particularly strong since the results from international comparative studies such as PISA and TIMSS, in science, reading and mathematics, caused concerns about Norwegian students' core competencies (Grønmo et al.; Grønmo & Onstad, 2009; Kjærnsli et al., 2007; Kjærnsli et al., 2004; Lie et al., 2001). The results from the large-scale assessment studies in Norway were confirmed in separate evaluation studies (Haug, 2003, 2004a, 2004b), particularly related to (1) teachers' unclear expectations of their students (Hertzberg, 2003; Klette, 2003) and (2) activities in the classroom lacking specific learning goals, which has been described as 'doing without learning' (Bachmann et al.; Klette & Lie, 2006).

Since the 1950s, UNESCO and the OECD have gathered information from school systems around the world, with the aim to compare students' knowledge levels (Stanat & Lüdtke, 2013). The challenge of measuring students' outcomes based upon comparable educational certificates across countries was soon recognised and this was the impetus

for eminent scholars[15] to meet from the fields of sociology, psychology, psychometrics and education in UNESCO's Institute of Education in Hamburg in 1958. At that historic meeting, they discussed whether one could develop tests that could measure students' learning outcomes across countries (Gustafsson, 2012). The result of the meeting was the First International Mathematics Study, which we later came to know in the form of TIMSS, initiated by the IEA. Similarly, the number of participating countries in PIRLS has increased from 35 countries in 2001 to 45 in 2006 and a total of 350,000 students from 50 countries in 2011 (Mullis & Martin, 2013). Like TIMSS, the PIRLS study is directed by Mullis and Martin from the TIMSS and PIRLS International Study Center at Boston College in the US. The goal of these international large-scale studies was to offer information to the participating countries about their students' learning outcomes, and to compare it with other nations (Stanat & Lüdtke, 2008).

It has been argued that the OECD has been particularly good at engaging with policy-makers around the world through PISA, and that the researchers behind the IEA studies have not to the same degree been able to influence the different countries' education policy. Relatedly, it appears that PISA results capture the public imagination to a larger extent. Researchers in Austria experienced that the public attention on PIRLS 2006, which was published in November 2007, decreased as soon as the PISA results were published a month later (Suchan et al., 2012). German PIRLS researchers described similar experiences. The PISA studies from 2000 and 2003 attracted considerably more attention than the PIRLS studies of 2001 and 2006. Since the fourth graders in the PIRLS study performed much better than the 15-year-olds in the German PISA study, the media and the public interpreted the findings as problems with the secondary education system (Lenkeit et al., 2012). Similar effects were found in Hungary, where poor PISA results attracted more attention than the better PIRLS results (Balkanyi, 2012). It appears that PISA is now making a bigger impact upon policy, the media and research than other international tests due to the ways in which the tests are governed and the dissemination channels. PISA's Governing Body is composed of policy-makers, whilst the IEA-produced tests are governed by researchers (Olsen, 2005 – Table 5; Stanat & Lüdtke, 2013).

In this chapter, we describe the findings from our review on international tests and their impact on learning.

7.1 Systematic review method: international tests

Our general method for conducting the systematic reviews was described in section 3.2. Here, we explain the details specific to the systematic review of international tests.

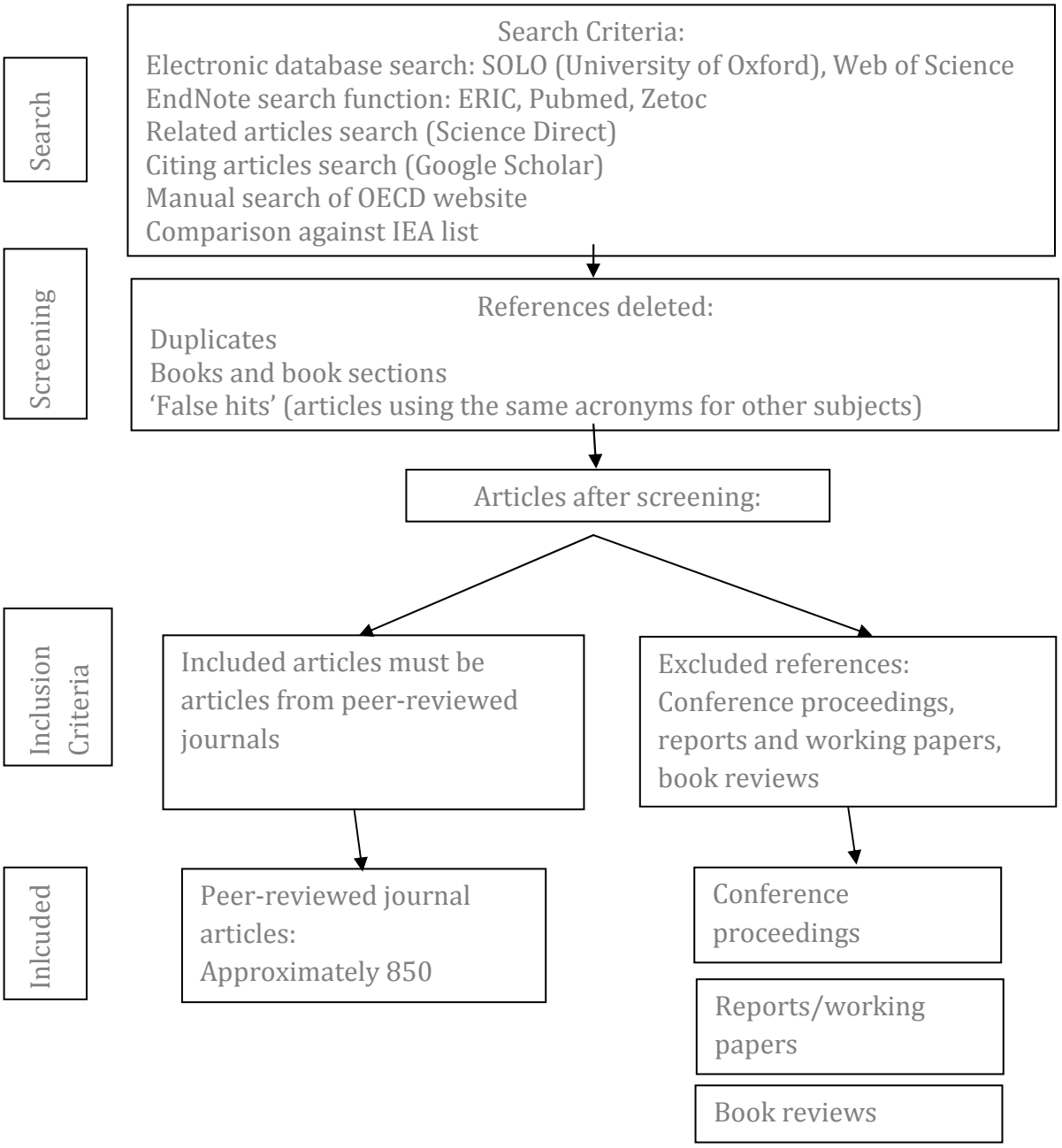
[15] Bill Wall from National Foundation for Educational Research, England, Arnold Anderson and Benjamin Bloom, University of Chicago, Robert Thorndike, Colombia University and Thorstein Husén, Stockholm University (Gustafsson, 2012).

Figure 6 summarises the process. We excluded studies on the National Assessment of Educational Progress (NAEP), even if American research and assessment policy has been highly influential on the educational discourse the last 100 years. Strictly, it is a national test, but it could be considered relevant due to its scale. An initial search detected more than 1,200 articles related to NAEP, so practical considerations also led to its exclusion. We did not include tests from South America and Africa, since these studies cannot be said to have influenced the Scandinavian context to the same degree as the OECD and IEA studies.

Only peer-reviewed articles published from 1990 to 2013 on international tests were included, as prior to this point, articles were more speculative in nature (eg calling for more systematic comparative studies). The search terms were: Third International Mathematics and Science Study (TIMSS), Programme in International Student Assessment (PISA), Progress in International Reading Literacy Study (PIRLS), Assessment of Higher Education Learning Outcomes (AHELO) and Program for the International Assessment of Adult Competencies (PIAAC). The search techniques were given in section 3.2. In total, approximately 200 articles were excluded based on the criteria. Due to the time lag between publication and articles appearing on Web of Science and other databases, the Kreiner and Christensen (2013) article was not identified by the search technique. We added this article to the database.

References were grouped according to the test, and articles that discussed more than one test were added to all relevant groups. More than 850 peer-reviewed articles on AHELO, PIAAC, PIRLS, PISA, and TIMSS were found. We used the Dominguez et al. (2012) article as a model article, but adapted the classification scheme outlined below. Articles were classified using the title, abstract and, where available, the PDF (approximately 75%). Where an article included an analysis with a number of variables, there was cross-classification. However, reading all 850 references in the time available was not feasible, so it is possible that references to some student-level variables (eg gender) are currently misclassified. A number of checks on the systematic search are planned, following use of the database for this review. Ultimately, the systematic review will also be published as an academic outcome of this grant.

Figure 6 Systematic search method: international tests



The following research questions were addressed regarding the focus of the article and these composed our classification scheme:

- Test influence upon research: which test(s) were the focus of the article?
- Discipline: which subject was the focus of the article (competencies, mathematics, reading, problem-solving or science)?
- Geographically impact: which country or jurisdiction was the focus? (A range of countries was sometimes coded.)
- Methodology: was the article a critical evaluation or empirical paper?

- Critique: which technical issues relating to the tests were investigated?
- Level: did the article focus upon the teacher (classroom), school and/or system levels?
- Student factors: which student-level data were researched?
- Test impact: articles discussing country performance, economic issues, the relationship between the tests and national tests, curriculum matters or influences upon government policies.

As discussed previously, our purpose here is not to present the systematic review, but to use it to support investigation of the relationship between the international testing literature and learning. We have selected two areas to focus upon here in this regard: the way in which learning is viewed through international tests and the impact of international tests upon learning. In a critical narrative evaluation, it is also incumbent upon us to outline the critiques of international tests that we find in the literature. We do so in each of these sections and include a third section that depicts the literature on technical debates, looking at whether the assessment technology delivers what it promises.

7.2 Approaches to learning in international tests

One of the main challenges in international tests is whether it is possible to compare results across different cultures, and whether it is possible to test in a reliable and valid way using the same tests across different contexts and cultures. The IEA studies have decided to use a *curriculum-oriented approach* and develop tasks based upon the participating countries' curricula. A valid test will then have items that match the international curriculum (Stanat & Lüdtke, 2013, p481). The OECD PISA study has developed what the OECD calls the *literacy-oriented approach,* where the test items are developed to measure what the 15-year-old students in participating countries are able to do at the end of their schooling, and how they are able to apply their knowledge in authentic situations, solving problems that they have not yet experienced (Nardi, 2008). This approach focuses more upon skills than content, although it is obviously very difficult to separate the two completely, conceptually or in practice. From a Norwegian perspective, the PISA approach has been described by some scholars to be a more modern approach for the future, since the main goal of the study is to measure 'what students are able to do in tomorrow's world' (OECD, 2012), while TIMSS, which is based upon the curriculum approach, tends to be seen as traditional, less able to change and not fit for the future (Østerud, 2006).

As an illustration of how these items might differ, it is worth looking at examples of released items from PISA and TIMSS in science. One example from grade 8 in a TIMSS

science task can illustrate the curricula approach.[16] The cognitive domain is knowledge, and the main topic *Energy Transformations, Heat and Temperature* in physics. The item has the headline *Molecules of liquid* when it cools and the question reads:

What happens to the molecules of a liquid when the liquid cools?

A) They slow down
B) They speed up
C) They decrease in number
D) They decrease in size

Students are supposed to select the correct answer, which in this case is response A. If we compare this item with an item from PISA, it is easy to see the differences. Figure 7 is part of an item in science, measuring students' ability to evaluate conclusions, and the theme is *Human Biology* in the area of *Science in life and health*.

To get full credit (two marks), the PISA rubric indicated that students have to write something like 'refers to the difference between the numbers of deaths per 100 deliveries) in both wards, due to the fact that the first ward had a high rate of women dying compared to women in the second ward, obviously show that it had nothing to do earthquakes. Not as many people died in ward 2 so an earthquake couldn't have occurred without causing the same number of deaths in each ward'. Partial credit is given to answers such as 'It would be unlikely to be caused by earthquakes because earthquakes don't occur frequently'. If students only state that earthquakes cannot cause the fever, they will not be rewarded any marks.[17]

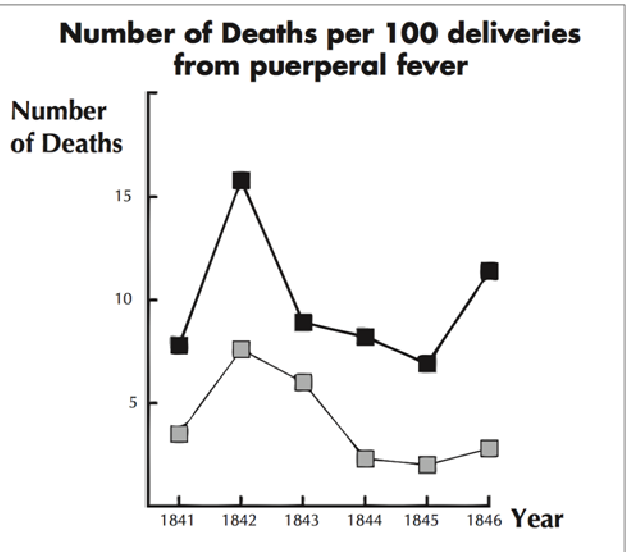
[16] The released items can be downloaded here: http://nces.ed.gov/timss/pdf/TIMSS2011_G8_Science.pdf.
[17] See the following website for released items and scoring: www.oecd.org/pisa/38709385.pdf.

Figure 7 PISA item in science

Semmelweis' Diary Text 1

'July 1846. Next week I will take up a position as "Herr Doktor" at the First Ward of the maternity clinic of the Vienna General Hospital. I was frightened when I heard about the percentage of patients who die in this clinic. This month not less than 36 of the 208 mothers died there, all from puerperal fever. Giving birth to a child is as dangerous as first-degree pneumonia.'

These lines from the diary of Ignaz Semmelweis (1818-1865) illustrate the devastating effects of puerperal fever, a contagious disease that killed many women after childbirth. Semmelweis collected data about the number of deaths from puerperal fever in both the First and the Second Wards (see diagram).



Physicians, among them Semmelweis, were completely in the dark about the cause of puerperal fever. Semmelweis' diary again:

'December 1846. Why do so many women die from this fever after giving birth without any problems? For centuries science has told us that it is an invisible epidemic that kills mothers. Causes may be changes in the air or some extraterrestrial influence or a movement of the earth itself, an earthquake.'

Nowadays not many people would consider extraterrestrial influence or an earthquake as possible causes of fever. We now know it has to do with hygienic conditions. But in the time Semmelweis lived, many people, even scientists, did! However, Semmelweis knew that it was unlikely that fever could be caused by extraterrestrial influence or an earthquake. He pointed at the data he collected (see diagram) and used this to try to persuade his colleagues.

QUESTION 1.1

Suppose you were Semmelweis. Give a reason (based on the data Semmelweis collected) why puerperal fever is unlikely to be caused by earthquakes.

.....

.....

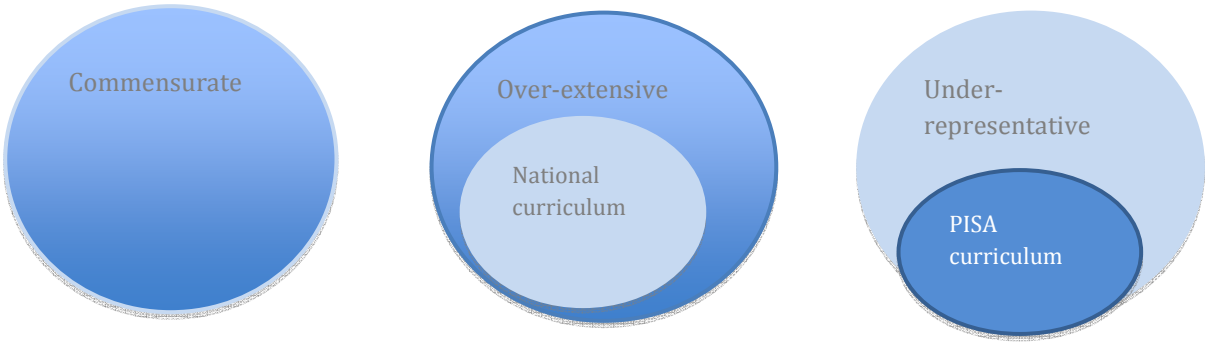
.....

As is evident just from looking at the items, the reading load in PISA items compared to TIMSS items is higher. TIMSS items tend to be isolated items, while PISA usually has what is called units within an item, such as with Semmelweis' Diary. First comes a text, graph or figure, then several units (questions), possibly followed by another text, followed by another unit (see Olsen (2005) – Table 5 for a fuller comparison between TIMSS and PISA). In addition, comparing the PISA items with the TIMSS items, students have to read and comprehend the text, interpret it, apply their knowledge and reflect

upon the answers in a different manner. Even though PISA has several multiple-choice tests, the overall reading load including open-response categories is a real difference between PISA and TIMSS. In other words, the literacy approach in PISA is focusing upon how students apply their knowledge in science, and it is expected that students can show what they have learned by the end of their schooling to solve future problems.

The use of international, 'curriculum-independent' tests to determine the quality and equality of national education system is not value-free. There are many potential problems with the globalisation and convergence agenda that could be associated with such an approach and researchers have called for research methodologies that describe supranational trends (Nóvoa & Yariv-Mashal, 2003, p430). Valuable aspects of a national education culture might not conform with the OECD's definitions of literacy, numeracy and scientific reasoning. For example, developing countries might emphasise health matters to a greater extent in their science curricula. Nardi (2008) outlined the culture-fairness issue well, describing three possible relationships between a PISA curriculum and a national curriculum (Figure 8). Where the OECD and national curricula are entirely commensurate, there should be no culture-fairness issue. However, in the case where the national curriculum is narrower than the PISA curriculum, students would be tested on material that they have simply not been taught. This is unfair to the individual students, but the tests are not high-stakes for them. For the country, the inference might be drawn that the education system is ineffective, but in fact there might be solid reasons for the design of the curriculum. For example, students might be required to study a larger number of subjects than in other countries; the national curriculum might emphasise a broad education. In the third situation, of under-representation of a national curriculum in the PISA curriculum, a country might draw the conclusion that it had not performed well when in fact its students knew a much broader range of material than was being tested. Thus, the test would not give students a chance to demonstrate what they know. We cannot interpret the results of PISA independently from the relationship with national curricula, so the claim for 'curriculum-independence' of the tests cannot be sustained.

Figure 8 Some possible relationships between PISA and national curricula



7.3 The impact of international tests upon learning

In a special issue of the *European Educational Research Journal*, Jenny Ozga wrote:

> Throughout the [European][18] countries [included in this research]… there is a growth in knowledge/information expressed in quantitative forms through comparative statistics, charts, diagrams and comparative tables, competency-based curricula and pedagogy that is partly driven by measurements of competence. There is a stronger emphasis on the production of knowledge about performance in quantitative forms that can be shared, analysed and translated into benchmarks and indicators. There is a growth of evaluation, accompanied by 'scientisation'. Accountability is clearly linked to the production and circulation of data-based knowledge and to introducing self-regulation (of teachers and pupils). There is increasing emphasis on technical, 'objective' knowledge as the basis for evaluation of performance by external agencies and internal audiences alike. (Ozga, 2012, p168)

Further, Ozga described how PISA has become a powerful policy instrument that penetrates different kinds of policy regimes in different countries, due to its high credibility and trustworthiness as 'the gold standard' (Carvalho & Costa, 2009). There is growing recognition of the power that such surveys have to affect educational and assessment policy (see, for example, Rinne, 2008; Grek, 2009; Grek, 2010; Lawn & Grek, 2012); the credibility that governments afford organisations like the OECD when responding to their own countries' PISA scores (Grek, 2009; Lingard & Rawolle, 2011); and the use of PISA scores as political tools to initiate educational reform and policy borrowing from one country to another (Bulle, 2011).

Alliances between organisations such as the OECD, the EU, UNESCO and the World Bank further serve to promote particular policy ideas and agendas. Rinne (2008, p676) wrote that

> …in the obscuring processes of the supranational homogenisation of education and educational policy and of national differences, supranational organisations, such as the OECD and the EU, play a significant role. It would also seem that the message, objectives and language of those organisations are cast in the same mould. They have started to speak in the same words with the same stress, repeating the same phrases about globalisation, economic efficiency and productivity, and swearing that globalisation is inevitable in the name of progress. In this discussion, the role of nation-states stays silenced in the background.

[18] Text in brackets added.

The OECD is independent of government policy and can therefore be cast as disconnected from national politics and superior in terms of its thorough, objective, technical, international evidence base. Lawn and Grek (2012, p134) argued that PISA results can be used as an 'externalisation' device by policy-makers within countries – as an appeal to a superior, international form of knowledge and standards.

Given this literature, we would expect to see policy convergence across jurisdictions as they implement the lessons learned from PISA. Holzinger and Knill (2005, p780) outlined five causal mechanisms of policy convergence: imposition (political pressure), international centralising (legal obligation), regulatory competition, transnational communication and independent problem-solving. Legal pressures are not likely to affect implementation of educational reforms based upon PISA, but we might expect the other four causes to come into play.

Interestingly, PISA appears to be used as a justification for educational reform even in countries where the results are good. In Finland, for example, the core curriculum was emphasised to a greater extent (Välijärvie et al., 2003, p51). In fact, the relationship between international test results and the particular reform introduced in countries is unclear. It is possible that, at least in some instances, international test results are used as rhetorical devices when they support the policy direction of politicians and policy-makers. Additionally, reactions to similar results differ across countries. For example, Greger (2012) reported that Germany was in shock about its PISA results and there was a large reaction, whereas similar results in the Czech Republic produced little policy response. Japan experienced a PISA shock since it had performed less well than on TIMSS (Westbury, 1992; LeTendre, 1999). Macnab (2000) analysed responses to TIMSS across 23 different countries and concluded that there were huge differences in the attention given to the results. Schwippert and Lenkeit (2012) also summarised the impact of PIRLS in 12 countries. They outlined the effects upon the quality of instruction, curricula, structure of the education system, teacher recommendations, resources, quality agencies, monitoring and assessment, reading programmes, programmes for special groups and standards (see p245–8). Five countries developed standards aligned with the PIRLS standards (Austria, Germany, Netherlands, New Zealand and South Africa).

Gür et al. (2012) also concluded that the Turkish government made use of PISA results to justify curriculum reforms that it had already decided to implement. Takayama (2008) identified this pattern with respect to reforms in Japan involving child-centred pedagogy and slimmed-down curricula. Rautalin and Alasuutari (2008) similarly reported that Finnish policy-makers explained successes with respect to their actions and any limitations were ascribed to external causes. In Portugal, Afonso and Costa (2009) gave a long list of pre-existing policy measures that were legitimised by reference to PISA data: from changing the teacher qualification framework to the

refurbishment of school buildings. The evidence for this approach to the use of PISA data by policy-makers is mounting. Indeed, there is evidence of politicians *creating* pressure by exaggerating the weakness of national PISA performance. Gür et al. (2012) termed this 'public scandalization'. Similarly, Lawn and Grek (2012, p134) wrote that

> …policy actors are using PISA as a form of policy legitimation, or as a means of defusing discussion by presenting policy as based on robust evidence.

They conclude that PISA was being used as a *governing device* at both national and international levels. So far we have been discussing the impact of international tests on educational policy broadly, as the literature reflects this. However, in this review the focus is upon learning specifically.

Whether the TIMSS approach to learning (content) or the PISA approach to learning (skills) is influencing education and assessment systems around the world could therefore have an impact on what kind of learning is seen as most valued. There is documented evidence that countries change their assessment and exam systems to align them with PISA tests. So far, little is known about whether this impacts directly upon students' learning.

Previously, it was documented that PISA had had a strong impact on the assessment systems in Canada, Sweden, Austria, Germany, Hungary, Ireland, Japan, Luxembourg, Norway, Poland and the Slovak Republic, (Breakspear, 2012; Baird et al., 2011; Grek, 2009; Takayama 2008, 2010; Ozga, 2012). In Norway, national tests in reading were developed using the reading framework from PISA, including the task format and suggested analysis of the items (Frønes et al., 2012), while Denmark has introduced national tests after low performance in PISA (Egelund, 2008). In Japan, one of the high-performing countries on PISA, analysis showed that the Japanese students performed better on tasks demanding memorisation than thinking skills, and as a result, Japan changed its national tests to more open-response tasks, like those in PISA (Schleicher, 2009).

Korea also introduced tasks that are framed like the PISA tasks in its university entrance exams (Schleicher, 2009). Puk (1999) cautioned against using TIMSS results to devise science curriculum policy across states in Canada, arguing that the curricular values differed. We anticipate more of this in the coming decades; due to globalising influences it is likely that there will be a greater degree of policy convergence internationally on what counts as valuable learning.

7.4 Technical debates in international testing

Many articles conducted secondary analyses of the international test data without questioning the nature of the underlying data. However, in addition to the conceptual debates presented above, there were a large number of articles researching the nature of the data underlying the tests. Given the growing importance of the international tests for policy, it is not only necessary, but healthy, that there is such body of work in the field.

We discuss below, some of the main themes, but there are many others not covered in this review: the lack of longitudinal data (eg Goldstein, 2004), response and test format familiarity (eg Elvers, 2010), motivation of the students taking the test (eg Eklöf, 2010), item readability (Dempster & Reddy, 2007), equating (eg Monseur & Berenzer, 2007) and issues relating to the questionnaires (eg Caro et al., 2013).

7.4.1 Sampling

To compare countries' educational profiles on international tests, representative samples of the learners at a specific age need to participate in the tests. After all, it is not interesting or very useful to know that 16-year-olds perform better in one country than 15-year-olds in another, or that more able samples of French students perform better than other countries' samples. To try to ensure this, guidelines are issued by the IEA and OECD to national testing centres. However, there have been some controversial issues.

Both PISA and TIMSS aim to collect data from high-quality samples in each country (n=4,000–5,000), following strict procedures for exclusion. They both started as pen-and-pencil tests, even though PISA now is gradually trying to become an electronic test. Both PISA and TIMSS use a rotation system of booklets, which makes it possible to test a larger number of items and give a more robust measure of each domain. What is not well known by the public is the fact that the rotation system means that students in a classroom taking the PISA test will be answering different items, depending upon which booklet they have. This is also the reason why some students might end up not answering some of the items in the domain in PISA, but still each individual student will be given a score in each domain of reading, science and mathematics. This score is based upon what the student would most likely have gained, based upon the score in the domain the student has solved items in. This is worth mentioning, since it is one of the controversial aspects of PISA (see section 7.4.2.3).

PISA uses a sample of students who are 15 in the year the test is taken. The grade therefore might vary from one country to another and a minimum of 150 schools are selected in each country (where this number existed) (OECD, 2013). In TIMSS, students are drawn from two populations: the fourth-grade population (the international target

population), which is all students enrolled in the grade that represents four years of schooling (population 1), and the eighth-grade student population, providing that the mean age at the time of testing is 13 and a half years old. See Martin and Mullis (2011) for further details.

While PISA samples 30 students from each school, TIMSS samples all students in a class. This is the main reason why it is not possible to link PISA data to individual teachers: students are sampled from the school level and might have several different teachers. In TIMSS on the other hand, it is possible to link student data to teacher data and TIMSS has a teacher questionnaire. Some argue that it is still not a wise thing to do, since students might have been taught by a good teacher the year before, or have several teachers, so it would be misleading to claim that the test scores on TIMSS could be explained by one specific teacher. PISA does not have a teacher questionnaire; instead it has a school leader questionnaire. Olsen (2005 – Table 5) argued that the different sampling designs between TIMSS and PISA can be seen as TIMSS having a multilevel modeling approach which to a larger extent targets factors in the classroom, while PISA targets issues related to schools.

The main difference between TIMSS and PISA is that TIMSS only measures science and mathematics, while PISA also measures reading in every cycle. In addition, PISA has introduced the domains of problem solving, financial literacy and electronic reading (see Frønes et al., 2011 for the Norwegian results in electronic reading).

In TIMSS and PIRLS the overall number of students excluded should not be more than 5% of the national target population. Exclusion rates are, in themselves, a technical issue often commented upon in the literature (eg PIRLs – Hilton, 2006). OECD (2013) gives the following rules for exclusions in PISA:

> **Exclusion codes in PISA**
> Code 1 Functional disability – student has a moderate to severe permanent physical disability.
> Code 2 Intellectual disability – student has a mental or emotional disability and has either been tested as cognitively delayed or is considered in the professional opinion of qualified staff to be cognitively delayed.
> Code 3 Limited assessment language proficiency – student is not a native speaker of any of the languages of the assessment in the country and has been resident in the country for less than one year.
> Code 4 Other reasons defined by the national centres and approved by the international centre.
> Code 5 No materials available in the language of instruction.
> Note: For a full explanation of the details in this table please refer to the PISA 2012 Technical Report (OECD, 2013).

The IEA uses the same exclusion rules for both PISA and PIRLS, and they are summarised as follows:

> **Exclusion rules for PIRLS AND TIMSS**
> Students with functional disabilities – these are students who have physical disabilities such that they cannot perform in the PIRLS and/or TIMSS testing situation. Students with functional disabilities who are able to perform should be included in the testing.

Students with intellectual disabilities – these are students who are considered, in the professional opinion of the school principal or by other qualified staff members, to have intellectual disabilities or who have been tested as such.[19]

7.4.1.1 Population

Some issues relating to the conceptualisation of the population of each country do not even arise in the academic literature. Some countries have a significant migrant worker population and domiciliary rights differ across countries. For example, in Singapore, until 2008, migrant workers' children were not permitted to attend state schools and were therefore not part of the population being considered. One third of the workforce is composed of migrant workers.[20] Differences between countries such as this are very important to the operation of their education systems. Countries that have seen significant rises in immigration, such as the UK, have to cope with the effects of different educational backgrounds, languages and cultures in classrooms. We also know from long-standing research as well as the results of international tests (eg Schnepf, 2007) that immigrant populations tend to do worse in the tests in many contexts (this pattern does not of course apply to all immigrant groups in all countries). Generally, though, excluding migrant workers from the population should have a very positive impact upon Singapore's test results and it does very well in the league tables. We may see an impact of the change to migrant workers' rights in Singapore upon future international test results, but this depends upon the extent to which it is feasible for these workers to access the education system for their children.

Similar issues arise in Shanghai, where admission to secondary school is almost impossible for children from rural, poor migrant families (Loveless, 2014). Access to school is controlled by the *hukou*, an identity card, which is used to manage where people live and access to services such as health and education. The university entrance examination (*gaokao*) can only be taken in the region of origin, so children are sent back to rural areas to study and be examined. The region of origin is inherited. As a proportion of the population, Shanghai had the lowest proportion of 15-year-olds of all

[19] http://timss.bc.edu/methods/pdf/TP_Sampling_Design.pdf.
[20] http://population.sg/resources/workforce-composition/#.Uw4aS_1HJFx.

of the jurisdictions participating in PISA (Loveless, 2014, p11, Table 1-1), a fact that is not explained by the one-child policy (Loveless, 2014, p12).

Countries' international test results depend upon factors well beyond the reach of international test sampling protocols. This example alone – ie definition of the population – serves to demonstrate the difficulties of conducting with confidence the comparisons supported by international test data across so many different countries.

At times, identification or processing of the population figures has also been in error. For example, Elvers (2010, p109) pointed out that 110% of all 15-year-olds in Germany and France were designated as enrolled in school by the OECD when presenting the PISA 2006 data. The same author pointed to anomalies in the data where countries were included in PISA even though their exclusion rates were higher than the 5% permissible (Canada, Denmark and Austria – Elvers, 2010, p110). Bonnet (2002, p396) points to discrepancies in the population sizes reported for schoolchildren in Korea.

For TIMSS, there questions have been raised about the proportions of students reported as still in school at higher ages, as this varies across countries. It has been noted too that many countries failed to meet quality control standards, at least in the early tests, which include sampling criteria (Bracey, 2000, p6–7).

7.4.1.2 Response rates

As mentioned above, certain exclusion criteria are allowed, as it is impossible to ensure that all of those sampled complete the test, but the return rates must be high. Some students might be sick on the day of the test and a small proportion of students have special needs that make the test unsuitable for them. Early PISA surveys had some significant problems with response rates. For example, Prais (2003) showed that England's response rate for the 2000 PISA test was only 60%, compared with over 95% in other European countries. Although Adams (2003) defended the sampling, there was subsequent debate and empirical work (Prais, 2004, 2007). The OECD later (2010, p1) recognised that the response rates for the UK were too low in both 2000 and 2003 and subsequent reports do not show trends using these data. Similarly, the Netherlands was excluded in 2000, but the US was not excluded even though its target response rates at school level were not reached in the first three cycles of the test (Elvers, 2010, p110–111).[21] More recently, there has been controversy over the response rate in Shanghai, which, it was claimed, represented 79% of the 15-year-old population.[21] This was already a low figure, but Andreas Schleicher (Deputy Director for Education and Skills, OECD) reported a lower figure of 73% to the UK parliamentary Education Committee.[22]

[21] The figure for the PISA response rate in the US is apparently now 89%. A figure of 79% is given for Shanghai in this *Education Week* article by Tucker and Schleicher, 2013:
http://blogs.edweek.org/edweek/top_performers/2013/12/response_to_the_brookings_institution_attack_on_pisa.html.

[22] UK parliamentary Education Committee, 5 March 2014, 0936-1129 Session, Wilson Room. Time stamp: 10:05:45. www.parliamentlive.tv/Main/Player.aspx?meetingId=15019andst=09:36:40.

7.4.1.3 Age

Another issue is that in some countries, students progress through the education system with their age cohort (eg Norway), whereas in others, students must pass a grade to move on with their age cohort (eg US). The sampling process for international tests has to choose either an age-based or grade-based process. PISA is an age-based sample of 15-year-olds and Wagemaker (2008, p273) showed that this means the comparisons are of different grades across countries. For example, in the PISA 2003 cycle, 99% of the sampled 15-year-olds in Norway, 61% in the US and 4% in Denmark were in the tenth grade. The distribution of students at this age across grades varied dramatically. Some countries, such as Germany and Switzerland, decided to add to the samples, so that they had grade-based information in addition to the country-level OECD comparisons (Prais, 2003, p146).

TIMSS has taken a grade-based sampling approach and has been criticised for having students take the 1998 test at different ages. Bracey (2000, p4) noted that there was a great deal of variability in the average age of students for the Final Year Study; it was 18.1 for the US, whereas it was as high as 21 in Iceland. Recently, Strietholt et al. (2013) analysed IEA studies of reading and concluded that the league tables for countries were invalid and trends were inaccurate, due to the differences in age and schooling in samples.

7.4.2 Psychometric models

7.4.2.1 Dimensionality

The international test results are treated as though there is a single underlying trait underlying the scores. Some articles question this assumption. For example, Goldstein (2004) showed that a two-factor model could be used for PISA mathematics data and Goldstein, Bonnet and Rocher (2007) showed that two factors fitted the PISA reading data better. Gustafsson and Rosen (2006) also found two factors each in the 1991 IEA reading study and the 2001 PIRLS study.

7.4.2.2 Number of parameters in the psychometric model

International tests use different psychometric models; PISA is based upon Rasch (1-parameter) and TIMSS uses a 2-parameter model. A number of researchers have investigated international test results and questionnaire data using different models (eg Cao & Stokes, 2008; Atar, 2011; Atar & Atkan, 2013). Other authors have used advanced quantitative techniques other than psychometrics to analyse the data (eg Hutchison et al., 2000; Goldstein et al., 2007; Frey & Seitz, 2011; Caro & Lenkeit, 2012). Goldstein (2004) queried the use of the Rasch model in PISA on a number of grounds, one of which was that when items do not fit this unidimensional model, they are discarded. Recently, Kreiner and Christensen (2013) conducted an analysis of the PISA reading scores and showed that a large number of items did not fit the model.

7.4.2.3 Plausible values

Data collection for the international tests is conducted using an incomplete design, in which students do not all take the same test items and the difficulty of the items is calibrated using psychometric models. Values for the items not tested for each student are imputed on the basis of the models; these are termed 'plausible values' (Wu, 2005 describes the process of creating plausible values). Controversy surrounds the use and creation of the plausible values because they can be difficult to recreate. Monseur and Adams (2009) investigated the implications of the single-level method used to create plausible values in PISA and showed that a hierarchical model was better for estimation of between- and within-school variances with plausible values.

7.4.3 Test translation

As the tests must be translated into the language of instruction in each country, there are naturally questions raised regarding the quality of the translation and its effects upon the scores. Grisay (2003) and Grisay et al. (2007) described the procedures used in PISA. Double-translation techniques are used, but some have questioned the time available to ensure a high-quality job is carried out. Some of these effects are difficult to research because the test booklets are not made public. A small selection of items is made publicly available in English and in French, but it would be much more transparent and foster higher quality research if at least half of the items were made available for each series in every language.[23]

Wiliam (2008) illustrated how translation effects can impact upon scores due to the meaning of words in different languages. In Welsh, students would be advantaged by use of the word 'speed' because a new term was introduced to denote it (*buanedd*), but disadvantaged by use of the word 'velocity' because a vernacular term was used (*cyflymder*). Thus, backtranslation cannot overcome all translation problems. Arffman (2010) similarly looked at translation effects for Finnish and Grisay (2003) discussed issues related to translation in French. As well as the meaning of words, there could be advantages and disadvantages cued by other linguistic factors, such as the frequency of words used, register or syntax (Solano-Flores, 2009; Elvers, 2010).

In the absence of the published test booklets, some authors have investigated potential translation effects by analysing the item scores across countries, after controlling for general performance on the test. Differential item functioning techniques have been used in a number of studies, with significant effects being demonstrated for TIMSS (Ercikan & Koh, 2005; Wu & Ercikan, 2006; Grisay & Monseur, 2007; Hauger & Sireci, 2008; Babiar, 2011; Mesic, 2012), PIRLS (Sandilands et al., 2013) and PISA (Grisay et al., 2009; Le, 2009; Oliveri & Ercikan, 2011).

[23] For example, see PISA 2012 released mathematics items: www.oecd.org/pisa/pisaproducts/pisa2012-2006-rel-items-maths-ENG.pdf.

Translation of the test booklets has resulted in different test lengths. Elvers (2010, p104) reported that German versions of the PISA test were 17% longer than the English version and Finnish versions were 8% longer, whilst the Irish version was 11% longer. Longer texts are more difficult for weak readers (Grisay, 2003). Ibrahim (2009) argued that the unique features of the Arabic language *per se* were an explanation for low reading scores in PISA in the Arabic test versions.

The language in which the test is taken is a complex matter, as there are many students for whom the majority language of the country is not their first language. There are indications that this issue is not always treated in a standard manner across countries (eg see the debate relating to PIRLS – Hilton, 2006; Whetton et al., 2007; Hilton, 2007).

The source language of the items is often English or French, but mainly English. For example, although PISA 2009 items were submitted by over 30 different countries in their national language, or in English, they were circulated to all countries in English for review (OECD, 2012, p82). This has been noted as a potential source of inequity by a number of authors (eg Elvers, 2010).

International tests have produced datasets that have been very fruitful sources of research in many substantive areas. We see from the above that there are numerous problems in interpreting their findings when the results are used for international comparisons of the quality of education systems. In the following chapter on validity, we note that test results are put to a range of purposes (section 8.2). Use of international tests for policy purposes at a national level is a consequential validity issue, so the discussions above about the sometimes-tenuous relationships between the test results and policy are fundamentally issues of validity.

Chapter 8 Validity and the cultivation of valuable learning

Paul Newton
Institute of Education, University of London

In an ideal world, each and every aspect of educational assessment practice would operate in harmony with the ultimate educational objective: the cultivation of valuable learning. Inevitably, though, the real world is far from ideal, and disharmonies abound. These disharmonies – these underlying tensions between assessment and learning – can be extremely challenging to apprehend and to understand, let alone to address. Apprehend and understand them we must, though, if we are to make progress towards improving synergy between assessment practices and the cultivation of valuable learning.

Questions of educational assessment practice – what counts as good practice and what counts as bad practice – are generally framed in terms of validity theory. Theories of validity have evolved over time and continue to evolve: partly for technical reasons, as the techniques and concepts of educational assessment have become increasingly sophisticated, but particularly for social reasons, as educational values have changed and educational contexts have become increasingly complex. In recent years, validity theory has been challenged from a variety of perspectives. Through a focused review of the major concerns that have been raised, we are able to shine light upon a range of uncomfortable tensions between present-day educational assessment practices and the cultivation of valuable learning.

The *Standards for Educational and Psychological Testing* (generally abbreviated to the *Standards*) is currently in its fifth edition, but is soon to be in it sixth. Since its inception, this document, which represents an official consensus statement of the North American educational and psychological measurement communities, has been very influential in shaping views on the sound and ethical use of tests, providing a basis for evaluating the quality of assessment practices. It has influenced conceptions of good practice in assessment internationally. The current edition introduces the concept of validity as 'the most fundamental consideration in developing and evaluating tests' (AERA et al., 1999, p9), which is a sentiment with which most assessment professionals would agree. Samuel Messick, one of the most important scholars of validity theory, went so far as to describe it as 'the only genuine imperative in testing' (Messick, 1989b, p11). The concept of validity is therefore fundamental in the sense of its signalling function: it focuses attention on what really matters when it comes to sound assessment practice.

The concept of validity has evolved significantly over time. It was originally defined as 'the degree to which a test or examination measures what it purports to measure' (Ruch, 1924, p13); and many still uphold this definition. Yet, even during the early years of validity theory, it was acknowledged that this definition was problematic for tests used primarily to predict, rather than to measure (see Cronbach, 1949). The first edition of the *Standards* fragmented the concept of validity into four different kinds: content, predictive, concurrent, and construct (APA et al., 1954). It took at least three decades before the concept of validity was reunified, under the influence of thinkers like Messick (1980, 1989a) who argued that all validity ought to be theorised along the lines of construct validity. For many, if not most, assessment professionals, validity now functions as an umbrella concept, intended to embrace a multiplicity of assessment concerns, including reliability (eg Messick, 1989a; Kane, 2006). This very brief account of the history of the concept sets the scene for an extended review of challenges to the modern conception, which is often referred to as the unified construct validity theory. A longer and more nuanced account is presented in Newton and Shaw (2014).

The central tenets of modern validity theory were epitomised in the official consensus statement that ushered assessment professionals into a new millennium of assessment practice (AERA et al., 1999). However, right from the outset, cracks were apparent in this new consensus, and they have continued to widen over the past decade or so. Through a focused review of these critical voices, we illustrate how modern thinking on validity and validation has been examined and found wanting. These perspectives shed new light upon the complex relationships between present-day educational assessment practices and the cultivation of valuable learning. They provide insight into challenges that must be tackled to increase synergy between assessment and learning.

8.1 Theorising validation

Although the principle underlying the reunification of validity has been generally accepted, the practical consequences are extremely challenging: modern validity theory is now far more tricky to understand than in previous decades; and modern validation practice ought now to be far more laborious. In the wake of Messick (1989a), the implication seems to be that any validation exercise ought to evaluate the plausibility of any intended interpretation of test scores in terms of logical analysis and empirical evidence garnered from as many sources as humanly possible. Not only does the scope of validation on this scale appear overwhelming, it is also unclear how to plan and undertake it. This is now widely acknowledged as one of the major challenges facing the assessment professions (eg Fremer, 2000).

In particular, many evaluators have struggled to implement recommendations in the 1999 *Standards* – for instance, Chapelle and colleagues, who found them very unhelpful when trying to evaluate a revision of the Educational Testing Service (ETS) Test of

English as a Foreign Language™ (Chapelle et al., 2008, 2010). The idea of validity as a unified concept implies that validation should be theorised as a major programme of scientific research. Unfortunately, although the 1999 *Standards* provided a clear list of ingredients for a research programme like this, it failed to provide a recipe for combining them. In recent years, leading scholars have claimed that argumentation theory holds the key to repairing this gap in validation theory.

The idea of validation-as-argumentation was originally proposed by Cronbach (1980a, b) although it has subsequently been developed by numerous measurement specialists, including Robert Mislevy (eg Mislevy, 2003a, b; Mislevy et al., 2003; Mislevy, 2009), Lyle Bachman (eg Bachman, 2005; Bachman & Palmer, 2010) and, in particular, Michael Kane (eg Kane, 1992, 2001, 2002, 2004, 2006, 2009, 2013). From this perspective, validation is tantamount to constructing an argument in support of the intended interpretation and use of assessment results. Ultimately, the strength of the argument determines the plausibility of the intended interpretation or the defensibility of the intended use. The conceptual tools of argumentation theory – framed in terms of claims, data, warrants, backing, and so on – help to scaffold the development and appraisal of the argument. As such, the real payoff from this approach is the structure and focus that it provides.

This was exactly what Chapelle and colleagues concluded when they switched from the guidance in the 1999 *Standards* to the argument-based approach, concluding that it helped them to specify the intended score interpretation, outline the essential research, locate research results within an overarching argument, further interrogate the argument, and ultimately to appraise its strength. In other words, it helped them to work out where to begin, how to proceed, and when to stop (Chapelle et al., 2010; Kane, 2013). Perhaps the most important claimed benefit of the argument-based approach is its potential to make the task of validation more tractable, by clarifying which aspects of the argument require most attention, thereby indicating where to focus scarce evaluation resources (Kane, 2013). The key concept, here, is the weak link in the argument chain, to which most attention ought to be paid.

In the context of high-stakes educational assessment, one of the most important weak links that the argument-based approach shines a light upon is the extrapolation from performance in testing situations (where a limited range of learning outcomes are assessed through more or less artificial tasks) to performance in the real world. The most obvious challenge is that no test could possibly sample all of the valued learning outcomes within a domain; and even test 'blueprints' are designed intentionally to over-simplify the complex sets of valued learning outcomes that comprise educational constructs, by operationalising them in terms of fairly crude content-by-process matrices. Having said that, fully representative sampling is not a necessary requirement for good assessment. As long as teachers still teach those aspects of the curriculum, and as long as students still attempt to achieve them, then there may be no problem: it may

still be legitimate to generalise and extrapolate from performance on the test to proficiency across the full complement of valued learning outcomes. Problems begin to emerge, however, when test design decisions preclude the assessment of certain valuable learning outcomes – typically those which are too hard, or too expensive, to assess accurately – under situations of high accountability when there are perverse incentives upon students and teachers to play the system. When there are incentives to not teach or acquire learning outcomes that are predictably absent from test forms, then conditions are ripe for score inflation, whereby students may genuinely perform better on tested material over time, without this reflecting an overall rise in attainment (eg Koretz, 2008, 2013).

Under circumstances like these it may no longer be legitimate to extrapolate from rising scores on a test, over time, to rising levels of proficiency across the full complement of valued learning outcomes. One possibility is that students are genuinely mastering the tested learning outcomes more effectively, but are no longer mastering the untested ones. A related possibility is that students are not even mastering the tested learning outcomes more effectively. Under the pressure of perverse incentives to improve test scores, teachers may decide increasingly to teach to the test rather than to teach for robust understanding (Shepard, 1997a). Training students in techniques for answering questions presented in predictable formats can help them to improve their test performances, but without any corresponding rise in their proficiency across the tested learning outcomes. Tests with the potential to support accurate measurement inferences under conditions of low stakes might therefore fail to support accurate measurement under conditions of high stakes.

Given considerations such as these, the holy grail of educational assessment would seem to be the development of tests worth teaching to (Popham, 1987; Resnick & Resnick, 1992) – that is, authentic assessments which represent the valued learning outcomes of a curriculum representatively through tasks that cannot be responded to effectively without deep understanding of the underlying concepts (Shepard, 1997a). In fact, precisely this ambition was embodied in the introduction of 'coursework' as a significant component of new General Certificate of Secondary Education (GCSE) examinations in the UK in 1988.

Unfortunately, despite principled proposals (eg Popham, 1987), the key to designing an assessment with the power to resist corruption by high stakes has not yet been discovered. In the UK, the coursework dream persisted for nearly 20 years, but eventually faded. In many subjects, what had begun as authentic assessment of 'the work of the course' was soon reduced to assessing performances on short formulaic tasks, for which students were coached. Allegations of widespread cheating proved to be the nail in the coffin of coursework (QCA, 2005; Hilborne, 2005; QCA, 2006). Coursework was replaced with 'controlled assessment' (Colwill, 2007; QCA, 2007). Controlled assessment tasks were designed only to assess learning outcomes that could

not be assessed by written examination, and were designed to increase levels of control over task design, administration and marking. The final nail in the coffin of controlled assessment came during 2012, when it became clear that teachers were playing the system in GCSE English speaking and listening assessments – using strategies such as marking borderline students unduly generously to the absolute limit of moderation tolerance – which undermined the integrity of the qualification (Stockford et al., 2012; Ofqual, 2012, 2013a, 2013b).

Returning to the argument-based approach, it is worth concluding with the observation that it is something of a double-edged sword. It is exactly what is required to deal with the complexity of modern validity theory, in the sense that it helps to shine light into the darker corners of validation research, by encouraging researchers to gather evidence that tests weak links rather than evidence which reinforces already strong ones. On the other hand, helping to structure this complexity and to take it seriously reveals the extent to which validation really does have the potential to become a huge undertaking, if it is to be done thoroughly. As more comprehensive argument-based evaluations begin to be published, this is becoming increasingly evident (eg Shaw & Crisp, 2012). The more thorough the ambition, the more complex and extensive the overarching evaluation argument becomes. This raises a fundamental challenge that is becoming increasingly apparent: as the concept of validity has been pushed to its limits – particularly in recognition that assessment practices cannot be evaluated in isolation but only in relation to broader concerns, such as teaching and learning – it has begun to take on epic proportions, and has left practitioners struggling to make sense of how to tackle it.

8.2 Multiple purposes

Not only has validity theory expanded in recent years, so too have the many uses to which educational assessment results are being put. When it is taken for granted that assessment results should be used for multiple purposes there is a requirement to evaluate them in relation to each of those purposes in turn. Once again, this renders the overall judgement concerning the acceptability of the assessment practice even more vast and complex.

This challenge arises when results from a single assessment procedure are used for more than one purpose. In a sense, given the preceding account of the evolution of validity theory, this is nothing new. For instance, Anastasi (1976) exemplified this challenge with reference to the use of results from a single mathematics test for four different purposes:

1. to test achievement in elementary school mathematics
2. to test aptitude for performance in high school mathematics
3. to diagnose learning disabilities
4. to measure logical reasoning.

Each of these different purposes implied a different inference from, or interpretation of, test scores – that is, in modern parlance, a different construct. In other words, if a test that was originally designed to measure one construct (ie achievement in maths) is subsequently used to measure another (eg logical reasoning), then there is a double duty of validation, in relation to each of the constructs. In this vein, Newton (2007a) distinguished the primary design-inference from secondary use-inferences, observing that the more dissimilar the two inferences were, the less likely it would be that a test designed to support one could also be used to support the other. Thus, results would inevitably be less accurate when interpreted as proxy measures of future attainment than when interpreted as direct measures of attainment at the time of testing. Newton (2007a) echoed a growing body of opinion (eg Resnick & Resnick, 1992; Pellegrino et al., 2001) in warning of the danger associated with the unreflective overuse of assessment data: an assessment that is fit for one purpose may be substantially less fit for another and might be entirely unfit for others.

In many countries around the world, it is not at all uncommon for results from educational assessments to be used for more than one purpose. Although this is not a new phenomenon, there seems to have been two important changes in recent years: first, results are now routinely used for a whole host of purposes; second, many of those purposes are premised upon the aggregation of results across students, classes and schools, adding new levels of complexity to the concepts, constructs and processes of validity and validation.

The numerous purposes to which assessment results are increasingly being put have been illustrated by various authors, including Baker (2000; 2013), Newton (2007a, b), Mansell (2007) and Ho (2012). Newton (2007a) identified 22 categories of decisions taken on the basis of assessment results, noting that even within each category there might be many different kinds of decision, implying that different assessment design decisions would be required to optimise results for each specific use of results. He subsequently observed that, by the end of the 2000s, results from national curriculum tests for 11-year-olds were being used to make decisions within at least 16 of these 22 categories (eg Newton, 2012b). This raised the obvious question of the legitimacy of each of these uses individually, which remains the subject of intense debate in England (eg House of Commons Children, Schools and Families Committee, 2008; Bew, 2011), but which has received very little attention in terms of systematic research. Baker (2013) similarly observed that developing validity arguments for entire sets of purposes is no small matter, and has rarely been done.

Developing this argument a stage further, Koch (2013) proposed that it is not simply a matter of investigating the validity of each use individually, because there is also potential for interaction between uses. She appropriated this concept of interaction to identify situations in which an activity or practice related to one use alters the validity of inferences related to another. When interaction occurs, she claimed, the argument-based approach becomes particularly challenging to apply.

An additional complication is that new purposes often emerge informally and gradually, and assert their dominance by stealth (eg Stobart, 2009). This phenomenon has been described as purpose drift (Ho, 2012) or purpose creep (Baker, 2013). Indeed, it is not unheard of for an assessment that has been designed to serve one purpose in one decade to come to serve a quite different purpose in the next. This clearly begs validity questions that, again, often remain unanswered.

Ho (2012) observed a particular risk associated with the fact that numbers 'travel' easily across levels of aggregation (from students to classrooms to schools, states, and countries) whereas the meaning of those numbers does not travel so easily. In fact, quite the reverse is true. He argued that modern validity theory has been reactive to purpose, whereas it needs now to be proactive: we need to anticipate known vectors of purpose drift to prevent the entrenchment of inappropriate uses of assessment results. To do this, he insisted, we need to develop models for predicting changes in score use over time (Ho, 2012.

The challenge of validating numbers at different levels of aggregation has been explored in some detail by Zumbo and Forer (eg Zumbo, 2009; Zumbo & Forer, 2011; Forer & Zumbo, 2011). Zumbo and Forer (2011) introduced the term 'multilevel measurement' to describe situations in which student-level assessment results are not even reported or used at the student level, being aggregated and used at higher levels (eg class, school, district, state, nation). They exemplified it using NAEP (US) and the Early Development Instrument (Canada). Underlying the concept of multilevel measurement is the concept of a multilevel construct: a construct that has potentially different meaning at the individual level versus the aggregate level. The crucial insight in multilevel measurement is that individual-level validity evidence (the basic currency of modern validation) may actually fail to identify invalidity at the aggregate level. Therefore, it is essential to ensure that the level of statistical analysis is aligned with the level of assessment interpretation and use.

Popham (eg 2007) identified a similar issue, which he claims affects all but a few of the standardised tests that are used for accountability purposes in the US: their instructional insensitivity. In other words, he claims that it is impossible to draw accurate inferences concerning the instructional quality of the teachers on the basis of aggregated results from these tests.

Misalignment between levels can be tricky to spot. Consider, for example, an alternative kind of multilevel measurement in which results from census tests (eg maths tests taken by all 11-year-olds) are reported and used at the student level, school level and at the national level. Statisticians are sometimes tempted to report the aggregate measures – ostensibly based upon results from entire populations of students – as though they were free from measurement error. Yet this begs the question of use and, therefore, interpretation. If school-level results are treated as measuring, say, school effectiveness, then the group of students in each class is simply a sample of all possible students who could have enrolled at the school, and the next year's students will be a different sample. More worrying still, the particular census test is actually just a sample (of size n=1) from the population of all possible measuring instruments, and the next year's test will be a different sample. In other words, it is essential to be explicit concerning measurement claims at all levels of measurement to develop and appraise appropriate evaluation arguments (eg Newton, 2009).

The challenge of multiple purposes is even more extreme when it involves results from multiple assessment procedures. Stobart (2001, 2009) faced this challenge when tasked with evaluating national curriculum assessment in England: a system comprising multiple, interacting assessment procedures – national tests alongside teacher assessments – designed to serve a range of purposes, with pupil-level formative feedback at one end of the spectrum and national-level progress monitoring at the other end. Stobart observed how managerial uses of results had come to dominate, detracting from the interpretation and use of teacher assessments. He concluded that the assessment system could only function effectively if the various components were kept in balance. He argued that synergy could best be achieved by drawing clearer distinctions between the components of the system and the purposes that they were intended to serve.

8.3 Both technical and ethical evaluation of consequences

Debate over the role of consequences in validity theory is not new, having gained momentum following the publication of watershed articles by Samuel Messick during the 1980s (eg Messick, 1980, 1981, 1988, 1989a, 1989b). In 1997, for instance, the journal *Educational Measurement: Issues and Practice* carried the first of two special issues on the topic, the editorial of the first being titled *The great validity debate* (Crocker, 1997). Unfortunately, this debate is still far from resolved, and leading theorists adopt radically different stances. Part of the reason for the lack of resolution is a lack of clarity over exactly what is at stake. In fact, there are at least three quite distinct debates, linked to the role that consequences play in:

1. the technical evaluation of intended interpretations and uses
2. the technical evaluation of intended policy impacts
3. the overall ethical evaluation of assessment policy.

The first issue is the most straightforward. It concerns the fairly uncontroversial, although poorly understood, claim that *post hoc* evidence from the consequences of assessing can provide important evidence concerning what the assessment actually measures and, therefore, whether it is suitable for making the kind of decisions that it has been used to make. Adopting a position closely aligned with Messick, Shepard (1997b) explained how evidence from the consequences of assessing was central to validating the intended interpretations and uses of assessment outcomes. Even theorists who see a less central role for this kind of evidence would generally accept its relevance to modern validity theory and validation practice.

The third issue highlights the role of evidence of consequences in judging whether it is socially acceptable to implement (or continue implementing) an assessment policy. On the one hand, it is obvious that consequences of assessing will have a major role to play here, and that this is ultimately an ethical evaluation concerning the trade-off between positive impacts and negative ones. Although this may seem fairly uncontroversial, the highly controversial aspect is whether this kind of analysis ought to be considered part of validity and validation. We will discuss the challenges arising from this question shortly.

The second issue highlights the role of evidence from consequences in judging whether it is possible to achieve a range of intended impacts through implementation of the assessment procedure or system. We turn to this challenge now.

8.3.1 Technical evaluation of intended policy impacts

One of the most important recognitions of recent years has been another gaping hole, this time in modern validity theory, where the technical evaluation of consequences ought to have been. Since the emergence of the field of educational and psychological measurement, it has always been recognised that assessments serve decision-making purposes. Yet it has only recently been recognised (by validity theorists, at least) that they also serve a subtly different kind of purpose too, and that this kind of purpose also needs to be part of the scientific evaluation process. Newton (2007a, b) distinguished between these two kinds of purpose as follows:

- at the decision level, the purpose of assessment is to enable a particular kind of decision; for example, students sit a school-leaving exam in science so that we can decide whether they have learned enough school science to justify their enrolment on a college science course

- at the impact level, the purpose of assessment is to bring about a particular kind of educational or social consequence; for example, students must sit an exam in science to 'force' them to learn the subject properly, and to 'force' their teachers to align their teaching of science with the national curriculum.

In other words, there are not two but three different evaluation questions, which concern whether it is possible to:

1. measure the desired attribute(s) using the assessment (measurement level)
2. make more accurate decisions by incorporating assessment outcomes into the decision-making process (decision level)
3. achieve a range of secondary impacts by implementing the assessment policy (impact level).

Notice that asking 'whether it is possible to achieve a range of secondary impacts' restricts the analysis to the realm of scientific evaluation. This is distinct from the kind of ethical evaluation of impacts and side effects mentioned above and described below. The technical evaluation of intended consequences involves specifying a mechanism by which each impact is supposed to be achieved – a theory of action – and then putting this theory to the test (see Hamp-Lyons, 1997; Lane et al., 1998; Baker & Linn, 2002; Forte Fast & Hebbler, 2004; Haertel & Herman, 2005; Nichols et al., 2009; Marion & Pellegrino, 2009; Bennett, 2010; Bennett et al., 2011).

In an article entitled 'How is testing supposed to improve schooling?', Haertel (2013) developed a distinction between direct actions (decision level) and indirect actions (impact level), observing that it was useful to divide the latter into incentive effects and messaging effects. Incentive effects are those related to the effort exerted by teachers and students to improve assessment outcomes, eg the use of testing to improve alignment between the official curriculum and the enacted curriculum. Messaging effects are more distant and concern the use of assessment outcomes to influence perceptions and beliefs, eg the use of assessment outcomes to bolster public perceptions that an education system is performing well or poorly.

Of course, there is nothing new in the idea of using assessment as a mechanism for achieving educational or social impacts. It has long been presumed that the 'threat' of a post-instructional test will motivate learners to achieve, although, in fact, this will not necessarily always be the case (Harlen & Deakin Crick, 2002). What has changed, as far as validity theory is concerned, is an increasing recognition that assessment cannot be evaluated in isolation, as though all that ultimately matters is the technical quality of measurement and decision-making. Assessment procedures have different kinds of purposes, and whenever it is claimed that a test can be used for a particular purpose, a burden of validation arises in consequence. If an important purpose of operating an assessment procedure is to improve teaching and learning, then the potential of the

mechanism to achieve this impact needs to be evaluated. In other words, educational assessment cannot be evaluated effectively in isolation, but only in broader contexts of teaching and learning.

The first and foremost challenge in evaluating intended impacts from assessment is that they tend to be insufficiently articulated (Haertel, 2013). The impact of testing upon motivation to learn provides a good example of an implicit justification for testing that is rarely made explicit and, therefore, rarely evaluated. Is it really true that students learn better under the 'threat' of a test? Does it make a difference whether they will be faced with an external test rather than an internal teacher assessment judgement? Do the stakes associated with the test make a difference, whether high or low? Are the impacts presumed to be positive for all students, or for only certain students? If there are differential effects, for what proportion of students are the effects positive? Putting questions like these to the test is an essential component of robust evaluation.

This broadening of the concept of validation to embrace intended impacts of assessment was made 'official' in the 1999 *Standards*. Unfortunately, even the *Standards* could have been clearer concerning the significance of consequential evidence, since it failed to distinguish clearly among the three key roles (identified above) that consequences might play in relation to validity (Newton, 2012a).

8.3.2 Ethical evaluation of assessment policy

From the scientific/technical evaluation perspective the focus is largely upon intended impacts and the plausibility of underlying theories of action. However, from the social/ethical evaluation perspective the focus is not simply upon intended positive impacts, but also upon negative impacts, anticipated or unanticipated, as well as upon unanticipated positive impacts. To do justice to genuinely ethical evaluation of educational assessment is one of the most significant of all the challenges we currently face.

As we discussed above, Samuel Messick championed the reunification that brought modern validity theory into being – we might call this the 'scientific transition', whereby construct validity became all of validity. However, at least in his earlier writings, he seemed also to have ambitions to champion an ethical transition – one that would have changed validity theory far more radically. As early as the mid-1960s, Messick distinguished clearly between:

> two major questions that arise in evaluating the appropriateness of a particular test administration:
> (a) Is the test any good as a measure of the characteristic it purports to assess?
> (b) Should the test be used for its present purpose?
>
> (Messick, 1965, p138)

> The first question is a scientific one; it may be answered by appraising the test's psychometric properties, especially its construct validity. The second question is an ethical one; it must be answered by evaluating the potential consequences of the testing in terms of human values. Both questions must be addressed whenever testing is considered.
>
> (Messick & Anderson, 1974, p22)

When Messick (1980) introduced his famous progressive matrix (reproduced in Figure 9) he represented these two questions within the first and second rows, respectively, and subdivided them further to distinguish between assessment score interpretation and assessment score use.

Figure 9 The progressive matrix

	Test Interpretation	**Test Use**
Evidential Basis	1 **Construct Validity**	3 Construct Validity & **Relevance/Utility**
Consequential Basis	2 Construct Validity & **Value Implications**	4 Construct Validity & Value Implications & Relevance/Utility & **Social Consequences**

(adapted from Newton & Shaw, 2014, p117)

By representing 'facets of test validity' in this manner (Messick, 1980, p1023), he provoked numerous theorists to respond that ethical evaluation ought to have no place in modern validity theory (eg Wiley, 1991; Maguire et al., 1994; Norris, 1995; Popham, 1997; Mehrens, 1997; Lees-Haley, 1996; Orton, 1998). Similar sentiments continue to be expressed in the present day (eg Borsboom et al, 2004; Lissitz & Samuelsen, 2007; Cizek, 2012). Other theorists, of course, have aligned themselves with the 'for' camp, and have welcomed the integration of ethical evaluation within validity theory (eg Cronbach, 1988; Linn, 1997; Moss, 1998; Brennan, 2013). Interestingly, there is evidence that Messick himself may have rolled back somewhat from his ambitions to champion an ethical transition, appearing to promote a far narrower concept of validity in his later papers, essentially restricted to the scientific evaluation of interpretations and uses (Newton & Shaw, 2014). The debate between the 'for' and 'against' camps is fairly easy to characterise. Those in favour claim that if ethical evaluation is not included within the concept of validity, then no one will take responsibility for it. Those against claim that if ethical evaluation is included within the concept of validity, then the concept of validity will collapse under its own weight and will be useless as a

framework for validation.

On one level, debate over the nature of the most important concept of educational and psychological measurement is, of course, profound. On another level, though, it is trivial. Cronbach (1988) put it like this:

> …validators have an obligation to review whether a practice has appropriate consequences for individuals and institutions, and especially to guard against adverse consequences (Messick, 1980). You (like Yalow and Popham, 1983) may prefer to exclude reflection on consequences from the meanings of the word *validation*, but you cannot deny the obligation.
>
> (Cronbach, 1988, p6)

In other words, whether you decide to include or exclude ethical evaluation from the concept of validity, you cannot deny the obligation to undertake it. Extending these sentiments, Newton and Shaw (2014) offered a revision of the original progressive matrix from Messick (1980) with an explicit intention to foreground ethical evaluation (Figure 10). However, rather than describing it as a validity framework, they described it simply as a framework for the evaluation of assessment policy.

Figure 10 A neo-Messickian framework for the evaluation of assessment policy

	Foci for the Evaluation of Assessment Policy		
	Mechanism for achieving the primary measurement objective(s)	**Mechanism for achieving the primary decision-making objective(s)**	**Mechanism for achieving secondary policy objectives**
Technical Evaluation (technical quality of mechanism)	**1** Is it possible to measure the desired attribute(s) using the test?	**2** Is it possible to make more accurate decisions by incorporating test scores into the decision-making process?	**3** Is it possible to achieve a range of secondary impacts by implementing the assessment policy?
Social Evaluation (social value of mechanism, expenses, payoffs, impacts, and side effects)	**4a** Have all of the primary measurement expenses been considered?	**4b** Have all of the primary decision-making expenses, payoffs, impacts and side effects been considered?	**4c** Have all of the secondary and incidental expenses, payoffs, impacts and side effects been considered?
	OJ **Is it acceptable to implement (or continue implementing) the assessment policy?**		

OJ = Overall Judgement

(adapted from Newton & Shaw, 2014, p191)

It is clear from Figure 10 that the overall integrated evaluative judgement with which the matrix culminates is an ethical one, concerning the acceptability of implementing (or continuing to implement) the assessment policy. It embraces the three relatively distinct technical evaluation concerns – measurement, decision-making and policy impacts – and also attempts to integrate each and every possible expense, payoff, impact or side effect.

This is where evaluation theory has the potential to accommodate the vast international literature on the educational and social consequences of assessing, which spans more than a century: from reports on the pros and cons of written examinations, published during the early years of the 20th century (eg Board of Education, 1911; Monroe & Souders, 1923), to the development of frameworks for analysing the range of impacts of assessment upon learning, developed towards the end of the twentieth century (eg Natriello, 1987; Crooks, 1988; Frederiksen & Collins, 1989), right through to the present day, with its emphasis upon cataloguing impacts which have arisen from the uses of assessment data for accountability purposes (eg Abu-Alhija, 2007; Mansell, 2007; Madaus et al., 2009; Berryhill et al., 2009; Hout & Elliott, 2011; Daly et al., 2012; Hamilton et al., 2012; Phelps, 2012; Polesel et al., 2012).

This, unfortunately, is also where ethical evaluation becomes extraordinarily challenging. For a start, there is no way on earth that every possible expense, payoff, impact or side effect could ever be catalogued. Logically, this would take us into the territory of the butterfly effect: *Does the flap of a butterfly's wings in Brazil set off a tornado in Texas?* (Lorenz, 1972). Even restricting the evaluation to more obvious and more critical impacts and side effects would present a cataloguing challenge of an almost unimaginable scale. To the extent that this evaluation needs to be prospective, and not simply retrospective, the job will be harder still, and will need to draw upon theories of action developed in cell 3 of the evaluation matrix.

Beyond the cataloguing challenge lies another challenge to which Messick drew attention: the fact that different stakeholders will value impacts and side effects differently. In other words, a positive impact for one stakeholder may be a negative side effect for another. Newton and Shaw (2014) attempted to navigate this challenge by arguing that the primary analysis ought to be conducted from the perspective of the assessment policy owner, ie the person with ultimate responsibility for the decision to implement the policy. However, to do so without at least representing other stakeholders' alternative perspectives to the policy owner would be remiss. So, once again, there is no escaping the obligation to embrace a wide range of stakeholder value-bases within the evaluation process, no matter how mind-numbing a challenge it might seem. Reflecting on the scale of challenges faced in attempting to do justice to ethical evaluation, Newton and Shaw (2014) observed that it is always better to make some crude attempt to approximate a systematic evaluation than not to bother at all. After all, policy decisions will continue to be made, and even an impoverished attempt at systematic evaluation would furnish better evidence than the anecdotal variety that would inevitably be relied upon in its absence.

The idea of different stakeholder perspectives and value-bases has been developed by a number of validity scholars over the past decade or so. McNamara (2001), for instance, identified potential tensions between three groups of language assessment stakeholders, each requiring different, conflicting qualities from assessment practices:

1. national and state policy makers and system managers – the most overtly powerful group – demanding policy relevance for accountability
2. language test developers – demanding clear interpretability, reliability and comparability
3. classroom teachers and learners – the least overtly powerful group – whose demands for meaningfulness in instructional process, facilitation of learning, enhanced quality of teaching, and minimisation of burden were only partly being met.

Developing this argument, McNamara (2006) identified two new perspectives within the international language testing literature, both reflecting an increase in social and political awareness: ethical language testing – representing the desire to make assessment practices more explicitly ethical, ie socially and educationally valuable; and critical language testing – representing the desire to make transparent the values underlying assessment procedures, as instruments of power and control, eg enhanced reflexivity and explicit development of alternatives to testing.

8.4 Philosophical presumption

One very significant element of the reappraisal of validity theory in recent years has concerned the philosophical foundations of the concept of validity. Joel Michell, for instance, has strongly challenged the presumption of many 'measurement professionals' within education and psychology that they are actually in the business of *measurement*. The idea that the attributes of educational assessment are measurable cannot, and should not, be taken for granted, he insists (eg Michell, 1999, 2000, 2009). As validity has traditionally been framed in terms of the extent to which it is plausible to interpret test scores as *measures* of specified attributes, this raises fundamental questions concerning the very enterprise of validation, as traditionally conceived. Others, including Michael Maraun, have taken the philosophical analysis even further, by arguing that the attributes of educational and psychological measurement simply are not measurable (eg Maraun, 1998). Yet others, including Denny Borsboom and colleagues, have insisted that, for the concept of validity to have any purchase at all, it needs to be grounded in a strong conception of measurement and an ontological commitment to the reality of the attributes which are supposedly being measured (eg Borsboom et al., 2004; Borsboom, 2005; Borsboom & Mellenbergh, 2007; Borsboom et al., 2009). Borsboom, in particular, has promoted a no-nonsense, back-to-basics reappraisal of validity theory. Not only does he reject the proposal to include ethical concerns within the concept of validity, he also rejects talk of validating interpretations: that is, we should return to talk of test validity, not interpretation validity.

These arcane discussions raise extraordinarily complex questions concerning the

nature of educational assessment. When we routinely refer to the products of learning that we claim to assess – to skills, understanding, abilities, concepts, aptitudes, and such like – what *kinds* of things are we talking about? How are they constituted? How can we know when to attribute, or not to attribute, them to students? When we routinely claim to measure these attributes, what kind of measurement are we talking about? Does the concept of educational measurement have any real meaning at all?

Although undoubtedly these are arcane questions, the answers that we provide to them, whether explicitly or implicitly, are far from inconsequential. A fairly strong sense of measurement, for instance, is at least implicitly embodied in traditional practices of ranking and grading students, in terms of the outcomes of their learning. This carries with it an air of 'scientific' credibility that may help to perpetuate perceptions of accuracy that are far higher than they ought to be (eg Newton, 2005). But maybe the very idea of ranking and grading students in terms of their learning outcomes is too radical an oversimplification? Maybe the idea of measuring students in terms of traditional one-dimensional, content-specified attributes/constructs (eg attainment in art, English, physics, maths, etc.) does more harm than good? Maybe the way in which we design between-learner discrimination into our educational assessments creates impressions of distances between learners that are largely unjustified (eg Wilson, 1998)? Recent debates concerning the philosophical foundations of the concept of validity have done more to divide assessment professionals than to achieve consensus. But, if nothing else, they provide an impetus to re-examine the many presumptions that are implicit in extant assessment practices.

8.5 Unstandardised assessment and decision-making

Perhaps the most profound critique of modern validity theory in recent years has concerned how to evaluate quality when considering assessments that are not overtly circumscribed by procedural specifications. The concept of validity was invented to deal with standardised assessments: not simply objective tests, but any kind of assessment that is defined in terms of a procedure. From this perspective, even a practical driving test involves a procedure, for instance:

- while there is considerable scope for variation in how the assessment is conducted (eg the driver will choose which car to drive, the examiner will choose where and when the emergency stop is to be performed);
- the procedural variation will be within specified bounds (eg there will be clear restrictions on what counts as a car for the purpose of the test);
- certain tasks will be specified (eg emergency stop, hill start);

- the processes for evaluating performances, and for aggregating evaluations to reach an overall outcome, will be specified (ie pass or fail, contingent upon 15 or fewer driving faults and no serious or dangerous faults); and
- the outcome will be linked to an intended interpretation (eg a pass implies that the driver is able to drive safely on public roads).

From this example, it should be clear that assessment procedures vary along a continuum, from tightly standardised to loosely standardised. But, as long as there is an element of standardisation, there is an assessment procedure; and it is the procedure that is put into practice when the assessment is administered (or, to put it more technically, it is the procedure that is replicated). Ultimately, from the perspective of modern validity theory, the procedure is the object of validation. That is, we are interested in evaluating the potential for drawing accurate measurement inferences on the basis of the assessment procedure, assuming that relevant administration conditions have been followed. This is related to Cronbach's observation that 'One validates, not a test, but an interpretation of outcomes from a measurement procedure' (Cronbach, 1971, p447). Because the interpretation of outcomes is the culmination of the assessment procedure, validating the interpretation is tantamount to validating the assessment procedure (Newton, 2012a). This analysis raises the very difficult question of how we might evaluate assessment and decision-making when lack of standardisation becomes extreme. Should we simply dismiss such practices as inherently unworthy of validation? Or should we seek to evaluate them from alternative perspectives? Over the past couple of decades, Pamela Moss has attempted to do the latter: first, in relation to educational assessment; second, in relation to decision-making on the basis of educational assessment outcomes.

During the 1990s, Moss focused primarily on traditional assessment procedures, albeit ones closer to the loosely standardised end of the continuum. For instance, an early paper identified a problematic tension between the standardisation of an assessment of critical thinking and its instructional relevance: using the same questions across a range of schools improved score comparability but undermined the provision of diagnostic information of relevance to the curriculum as manifest in each classroom (Moss and Koziol, 1991). In 1992, she traced evolving conceptions of validity within the mainstream tradition, beginning to question their epistemological principles, their privileging of standardised assessment, and their potentially negative impacts as mechanisms for exercising power over educational practices (Moss, 1992), echoing sentiments from Cherryholmes (1988). She subsequently developed this critique with the almost heretical claim that reliability is not necessarily an essential ingredient of sound assessment practice (Moss, 1994). The basis for her claim, once again, was that privileging standardisation – to ensure reliability of results – was potentially in tension with the development of good teaching and learning in schools, given that there are 'certain intellectual activities that standardized assessments can neither document nor promote' (p6). She introduced the idea of 'more hermeneutic approaches to assessment'

(p7), which she characterised as those towards the loosely standardised end of the continuum. From a hermeneutic perspective, she proposed, validity is maximised through dialogue concerning alternative interpretations of authentic, contextualised (unstandardised) evidence, rather than by foreclosing dialogue through the privileging of inauthentic, decontextualised (standardised) evidence.

Through a steady stream of publications spanning more than a decade, Moss has developed her argument (eg Moss, 1995, 1996, 2003, 2013; Moss, Pullin, Gee & Haertel, 2005; Moss et al., 2006). Her intention has not been to overthrow modern validity theory or traditional assessment practice but simply to highlight limitations and to legitimise alternatives. Central to her promotion of interpretive perspectives – from hermeneutics, to phenomenology, to critical theory and sociocultural theory – has been a desire to promote an understanding of meaning in real-world contexts of practice. She has presented her position as an invitation to engage with a range of epistemological questions:

> Is it more valid to have exercises evaluated independently, decontextualized from the rest of the performances? Is it more valid to preclude debate among readers about the actual performances in question? Is it more valid to evaluate performances isolated from the everyday context in which they were produced? Is it more valid to match cases to pre-existing interpretations than to allow the interpretation to be constructed in light of the cases? We have a long history that implies the answer to these questions is yes…There are traditions with equally long histories that suggest the answer to these questions is no.
>
> (Moss, 1996, p26)

This century, Moss has focused particularly upon the challenge of developing a robust validity framework to guide thinking and action in relation to classroom assessments, most notably on-the-fly formative assessment interactions. These tend to be about as loose as it is possible to be, in terms of the standardisation continuum, if it is even meaningful to consider them standardised at all. Indeed, the focus is primarily upon an assessment interaction, rather than an assessment procedure, with a focus on the entire social situation, ie student performance in context (Moss, 2003). At a higher level of abstraction the focus might be upon general principles for effective formative assessment practice, designed to help teachers to navigate, elicit, assimilate, integrate and successfully use the multiple, interacting sources of evidence that emerge from the complex social environments within which they practice, as well as upon the resources – material, conceptual, organisational, cultural, etc. – that allow those principles to be applied effectively (Moss et al., 2006; see also Crooks, 2011).

Admittedly, from this perspective, in the absence of a clearly defined assessment procedure, it is actually quite unclear what ought to constitute the appropriate level of,

and focus for, evaluation. Sometimes, the suggestion seems to be that each and every actual, local, situated interpretation or use of an assessment outcome is a potential candidate for validation research. Other times, higher levels seem to be implied. To be fair, similar ambiguities can also be found in the mainstream literature. For instance, when Mislevy illustrates his argument-based approach to validation (eg Mislevy, 2009), he tends to frame it in terms of the accuracy of actual (specific) interpretations of assessment evidence; cf. Kane (eg Kane, 2006), who tends to illustrate his argument-based approach in terms of the intended (general) interpretation of assessment evidence. Even Cronbach (1971) seemed a little unclear as to whether he considered the focus of validation to be actual specific interpretations, intended general ones, or both.

An interesting question is whether, from the perspective of formative assessment, validity is more about learning than about assessment. On the assumption that a defining feature of formative assessment is the requirement to close the gap between where students currently are in their learning and where they need to go (Black & Wiliam, 1998a; Assessment Reform Group, 2002; Gardner, 2012) a number of writers have suggested that validity for formative assessment ought to be theorised primarily as a matter of consequence, ie assessment-contingent learning (Wiliam & Black, 1996; Moss, 2003; Stobart, 2006). Other writers have adopted a different position, dividing formative assessment into discrete phases, eg assessment *versus* instruction, and have suggested evaluating these phases in relation to distinct criteria (Nichols et al., 2009).

Recently, Moss and colleagues have broadened their critique of modern validity theory to encompass challenges faced by local users of assessment data, eg educational administrators and other decision-makers who are required routinely to integrate multiple lines of assessment evidence in the course of their everyday professional practice (eg Moss et al., 2006; Moss, 2013). Moss et al. (2006) noted how the *Standards* (AERA et al., 1999) insisted that important decisions about students should not be based on single test scores, but failed to explain clearly how multiple lines of evidence ought to be integrated defensibly, particularly in the localised, situated, dynamic and partially unique contexts that characterise everyday decision-making. In contexts like these, Moss et al. (2006) argued, hermeneutic and sociocultural perspectives offer important insights into validity and validation, enabling the complexity of assessment and decision-making to be effectively theorised, guided by the notion of question-driven inquiry (Moss, 2013). This is to recommend new research agendas which focus on evidence-informed interactions, informational needs, community norms and routines, social and structural supports and constraints, points of developmental leverage, and so on (Moss, 2013).

DeLuca and Koch have followed in this tradition, arguing that the complexity which characterises educational assessment in the 21st century – extremely high stakes, increasingly complex constructs, multiple interacting uses of results, multiple stakeholders – demands new ways of thinking about validity and new approaches to

validation (DeLuca, 2011; Koch & DeLuca, 2012). Essential, they believe, are approaches that prioritise depth, perspective and values. They have promoted narrative case description and an interpretive approach to validation in which validity is viewed as socially and temporally situated and multifaceted. Like many critics of modern validity theory, they see an urgent need to provide more genuine, authentic and subtle accounts of educational assessment practices.

8.6 Tensions between assessment and learning

Over the past century, the concept of validity – an idea invented to focus attention on what really matters when it comes to sound assessment practice – has been constructed, deconstructed and reconstructed in response to technological advances, but also in response to social and cultural change, including changing values and increasing recognition of the complexity of the contexts within which assessment practices operate. This ongoing critique and, in particular, the critique of modern validity theory over the past couple of decades, has the potential to illuminate many of these underlying tensions between assessment and learning, as a necessary precursor to policy decisions which might help to improve synergy between them. The foregoing review has brought many such issues into the open for closer scrutiny.

The challenge of theorising validation practice and its proposed resolution, argumentation theory, has helped to illuminate a particularly problematic 'elephant in the room', the extrapolation link: we all know this link is highly problematic, but we often tend to ignore it. Conversely, argumentation theory forces us to acknowledge the very fallacies that help us to avert our gaze from this elephant. Most important, in this respect, is the fallacy of gilding the lily: we standardise and standardise in order to improve reliability; then we seek out all sorts of evidence with which to investigate whether we have managed to achieve this goal; and the more evidence we collect, the happier we are that we have successfully fortified the generalisation link. But what about the extrapolation link? Are we really at liberty to extrapolate from performance in our (heavily standardised) test situation to messy, complicated, real-world settings?

Incidentally, the risks of perverse incentives within the testing profession are no different from the risks of perverse incentives in any other profession, when being held to account in terms of performance measures. That is, if test developers are held to account primarily in terms of aspects of quality that are easy to measure (eg evidence of generalisability) then we should not be surprised if they aim to maximise those aspects, even at the expense of other aspects of quality that may be more important but harder to measure (eg evidence of extrapolatability). Again, this is not simply an issue of measurement quality, because there are likely to also be consequent impacts upon learning quality, when the problems that students are taught to solve in the test

situation depart radically from the problems that they need to be able to solve in the real world.

The value of the argument-based approach lies in being designed to expose such practices, by directing attention to all links in the chain, indeed by focusing attention upon the weakest links. Embracing the argument-based approach is not plain sailing, of course. Exposing the scope of comprehensive evaluation is sobering, to say the least, and possibly even off-putting. Moreover, any comprehensive evaluation is likely to uncover the odd weak link, here or there; and remember that a single link is supposedly sufficient to undermine the entire chain. The argument-based approach quite rightly focuses evaluators on potentially bad news stories: it recommends that they should seek out evidence and analysis that is most likely to be problematic, to be able to judge whether it is problematic enough to undermine the link. This is good evaluation. But it is not comfortable evaluation, particularly when the evidence and analysis is being generated and published by the agency responsible for developing the test.

The challenge of multiple purposes reminds us that it is insufficient to evaluate the degree to which a test measures the attribute that it has been designed to measure, if it is also being used as a proxy measure for a whole host of other attributes. It is well known that a test designed to function optimally for one purpose will not necessarily function optimally for another and may be entirely useless for yet another. So it is remarkable how infrequently validation research programmes focus explicitly upon secondary use-inferences beyond the primary design-inference. Most importantly, the challenge of multiple purposes foregrounds the threat of purpose creep, particularly the phenomenon of purpose displacement, eg when the use of results to manage an educational system displaces the use of results to support classroom learning. We need better ways of apprehending, monitoring and evaluating trends like these, to ensure that problematic interactions are dealt with before practices become so entrenched that they are no longer malleable.

The challenge of technical evaluation strikes right at the nexus of assessment and learning. Ultimately, it shines a light on the fact that assessment systems serve different kinds of purpose, and that each kind of purpose needs to be evaluated, not simply those related directly to measurement and decision-making. It is important that these secondary objectives, particularly those related to positive impact upon teaching and learning, are now being evaluated alongside the primary measurement and decision-making objectives, for the simple reason that they are all necessarily part of the same overall cost–benefit argument for implementing the assessment system. What is particularly important is that this new imperative will begin to make explicit the many unexamined, implicit justifications that help to sustain assessment practices around the world. It seems quite possible that certain ostensibly plausible justifications may collapse when scrutinised under this new microscope.

The challenge of ethical evaluation also strikes directly at the nexus of assessment and learning, and also particularly in relation to the impacts of assessment upon teaching and learning. The ethical perspective is broader, though, attempting to embrace the full range of impacts, from positive-anticipated to negative-unanticipated. Moreover, it is concerned explicitly with the social value of these impacts. This is the ultimate currency of the overall cost–benefit argument for implementing the system. With this challenge more than with any other we are struck by its enormity: the vast catalogue of potential impacts, as well as the different values that a multitude of stakeholders will bring to the table. However, only by taking this challenge seriously can we move educational policy-making beyond the realm of anecdote that is so easily abused by rhetorically gifted policy-makers. In a sense, the challenge from this perspective is how to see the wood for the trees. It has become clearer how educational systems are populated by many different kinds of actor, each with their own interests and motivations, and each making potentially conflicting demands of those systems. If it is meaningful to speak of the cultivation of valuable learning as the ultimate educational objective – which we believe – then how can all of those potentially conflicting demands be traded off in a manner than optimises and maximises synergy between assessment and learning?

The challenge of philosophical presumption has illuminated some of the darkest areas of assessment practice. To raise this challenge is to ask the most fundamental questions of all: what could it mean for a student to have achieved (or not have achieved) certain learning outcomes from a course of study? How could we know that they had achieved (or not achieved) them? To practice assessment is to presume, or perhaps to construct, answers to these questions, whether we realise it or not. Yet whether we presume, or construct, the best answers to these questions is a moot point, and one that is always important to return to. By doing so, we open horizons of possibility that would otherwise remain unexplored.

Finally, the challenge of unstandardised assessment forces us to appreciate that a very large amount of everyday, real-life, assessment practice and decision-making lies beyond the scope of modern validity theory, owing to its authenticity, that is, its complexity and consequent lack of standardisation. In the same way that we are beginning to get to grips with the technical and ethical evaluation of consequences, we also need to get to grips with the evaluation of unstandardised assessment and decision-making. Only by doing so can we hope to understand what might count as technical quality and social value in relation to the many educational practices which occur in localised, situated, dynamic and partially unique contexts – from a teacher's minute-to-minute instructional decisions to a senior manager's annual performance-related pay-award decisions. Only by understanding these contexts will we be able to identify and consider how to resolve the underlying tensions between assessment and learning.

Chapter 9 Review of Norwegian doctoral dissertations on assessment, 1993–2014

Therese N. Hopfenbeck and Anna T. Steen-Utheim
OUCEA and BI Norwegian Business School

9.1 Introduction

Doctoral students form a large part of the future capacity of the Norwegian nation in terms of educational assessment research. Therefore, it is important to know about the knowledge and skills that have been developed from the doctoral research undertaken this century and what it has contributed to the research field. We map the work conducted over the past 20 years in the area of assessment and learning in Norway. As mentioned in the introduction to this review, educational research needs to be more cumulative in nature (Oancea, 2005; James, 2012). Doctoral research is not always published, or easily accessible, as we will see below. As such, it is important to document this body of work, so that other researchers can learn from it and build upon it. Further, this map tells us what fields of enquiry have been given most attention at doctoral level within Norwegian society, which in turn tells us about the current state of funding and research interests.

Non-completion of dissertations has been identified as a major problem in many countries including Norway, but research in Norway showed that the completion rate and time-to-degree in doctoral training improved considerably, from 30% completing in five years in 1980 to 60% completing in five years in 2011 (Kyvik & Olsen, 2013). While many countries do not offer funding for all PhD students, two thirds of all doctoral students in Norway have temporary posts for three of four years, with a salary similar to a graduate-level position in the public sector (Thune et al., 2012).

The number of dissertations in Norway has also increased. In a country of only 5 million people, there were 9,000 doctoral students in 2011, most of whom were in medicine and the health sciences; fewer were in the humanities and social sciences such as education. Even though Finland still has the most doctoral students of the Nordic countries, Norway and Denmark have now increased the numbers. From 2002 until 2011 a total of 5,790 candidates were given the doctoral degree in Norway (Thune et al., 2012).

The rise in the number of dissertations is due to funding, support in research teams, and the development of doctoral training programmes, amongst other things. The

proportion of candidates who are female also increased, with a rise from 10% of the candidates in 1980 to 46% in 2011. The proportion of candidates who are foreign has also increased, from 5% in 1985 to 33% in 2011 (Kyvik & Olsen, 2013).

On behalf of the Ministry of Education, the Nordic Institute for Studies in Innovation, Research and Education (NIFU) registered candidates in the field of education who were awarded doctoral degrees in Norway from 2008 to spring 2011. The goals of the work were to: (1) present the state of knowledge in the field of education and to detect which areas needed to be further researched, (2) give an overview of institutions and subjects and (3) analyse the backgrounds of the candidates. All institutions in Norway were asked to participate. The institutions reported 119 doctoral degrees in the field of education, and of these 87 were women. The mean age was high compared to other sciences, 37 years old for men and 39 for women (Gunnes et al, 2011). The numbers of degrees were 33 in 2008, 35 in 2009, 36 in 2010 and 15 in spring 2011. Half of the dissertations related to elementary and secondary comprehensive schools, with the second biggest area being higher education, followed by dissertations on upper secondary schools. Compared with the total number of doctoral degrees awarded in Norway, eduation is a small but growing research field.

9.2 The defence

In comparison with other countries such as England, New Zealand and Spain, there have been few empirical studies investigating the evaluation of doctoral dissertations in Norway (Kyvik, 2012). The Norwegian Association of Higher Education Institutions has issued recommended guidelines for the PhD degree as follows:

> A doctoral thesis must be an independent piece of scientific research that meets international standards with regard to ethical requirements, academic level and methodology used in the research field. The thesis must contribute to the development of new knowledge and achieve a level meriting publication in the literature in the field.
>
> (UHR, 2011, cited by Thune et al., 2012)

The Norwegian assessment system of PhDs includes external reviewers on the examining committee, which is unlike other countries such as the US, where there are usually no external assessors on the committee. The supervisors in Norway are not part of the committee. A committee of three members assesses the dissertation and at least one of the members has to be from a foreign university. The outcome of this stage of the examination may be:

1) Approved: the thesis is found to be worthy of public defence
2) Revision of the thesis: the PhD candidate is asked to revise the thesis within three months

3) Rejected: the thesis is found not to be worthy of public defence.[24]

Unlike in Germany, the accepted dissertations are not given marks; they are either approved or not. The candidate will, if the dissertation is approved, publicly defend the thesis and be questioned by two members of the committee. First, the candidate will give an approved trial lecture on an assigned topic, and if the lecture is approved, the candidate will publicly defend his or her thesis; this is what is known as the 'Disputation'. Usually friends and family attend the defence, including colleagues from the higher education institution, and it is followed by a reception and formal dinner. The candidates presented in the current report have all defended their dissertations publicly, after having their dissertation approved to be worthy of public defence.

9.3 Method

To get an overview of published doctoral research on assessment and learning since 1993, a search of all higher education institutes in Norway was conducted.[25] Not all institutions offer PhD programmes, but the ones that do have been included in this review. For example, the University of Oslo offers PhD programmes in the following subjects: Education, Science Education, Learning and ICT, as well as a number of others.

As we are aware of the locations where research on assessment and learning and international tests is most prolific, we specifically targeted the University of Oslo, University of Bergen, Trondheim Norwegian University for Science and Technology (NTNU), University of Stavanger, University of Agder, University of Nordland and Akershus University College of Applied Sciences, to avoid missing any theses that may not have been listed on other systems. Further, we conducted searches for doctoral work on each of the institutions' websites. Some higher education institutions publish lists of published PhDs, and some have their own research engines (eg Bora, the database at the University of Bergen; Oda, the database at Oslo and Akershus University College).

In addition, we used the University of Oslo Library search engine BIBSYS (http://ask.bibsys.no/ask/action/stdsearch) and its open research archive, DUO. For the first search, we used the keywords 'assessment and learning', 'feedback' and 'formative assessment', in the English and Norwegian languages. Where the search engine allowed for defining the selected material type, we limited it to 'dissertations'. However, the search on BIBSYS only resulted in two hits. If the keywords 'assessment

[24] See the following link for *Guidelines for Assessing the PhD Degree* at HiOA Oslo and Akershus University College for Applied Sciences: www.hioa.no/eng/Studier/Lov-og-regelverk/Guidelines-for-Assessing-the-PhD-Degree-at-HiOA. Other examples are from the University of Bergen: http://regler.app.uib.no/regler_en/Part-2-Research-Education-Dissemination/2.1-Research/2.1.1-Administrasjon-av-forskerutdanning/REGULATIONS-FOR-THE-PHILOSOPHIAE-DOCTOR-PhD-DEGREE-AT-THE-UNIVERSITY-OF-BERGEN

[25] Lists of PhDs are published on www.uv.uio.no/isp/forskning/doktorgrad-karriere/fullforte-doktorgrader and www.ntnu.no/doktorgrader/phd from the University of Oslo and NTNU respectively.

and learning', 'feedback' or 'formative assessment' were not mentioned in the abstracts of the dissertations, these would not be included in the search results. As a consequence we have found several dissertations through our network, as explained below, with highly relevant content, but due to a lack of relevant keywords in their abstracts, they were not detected in our electronic searches.

We used our networks of assessment contacts to find additional doctoral work. This turned out to be the most effective strategy in collecting an overview of doctoral work. One of the reasons that some of the dissertations are not electronically available is the fact that they are published as articles, and the different journals will not allow candidates to publish their dissertations online for free access. Furthermore, those students who publish traditional monographs may not wish their work to be available electronically before they are able to publish in peer-reviewed journals. Several of the dissertations did not include an abstract, or keywords and it was therefore difficult to find the relevant material. We contacted authors with missing abstracts, and they have kindly submitted abstracts based upon their theses for this report and given their agreement to our publishing them here.

9.3.1 Limitations of the review of doctoral work in Norway

Since the search for Norwegian reviews was challenged by the fact that many of them were not available online, we cannot be sure that we have not missed any dissertations. Another limitation is of course the fact that we have not read all the dissertations. While most of the dissertations were between 200 and 300 pages, two of them were more than 600 pages long. We retrieved half of them and checked the content against the abstracts, and skimmed the dissertations with focus upon the research questions, method used, data and results.

9.4 Findings

Twenty-nine dissertations were identified, with 17 focusing upon assessment and learning and eight on international tests; two are on national tests and two are on national exams in Norway (Table 2). The Norwegian Directorate of Education and Training has previously expressed a concern about lack of assessment literacy amongst Norwegian teacher educators and researchers (Hopfenbeck et al., 2013). This volume of completed doctoral work is reassuring in the face of this concern.

The category of assessment and learning also includes dissertations that used the word 'evaluation', a word that historically has been used in the field of school-based evaluation research in Norway (see Haug & Monsen, 2002 for a review). We also uncovered doctoral work that is in the process of being submitted to a doctoral committee for review (five likely to be published in 2014). We are also aware of doctoral work in process that will most likely be published within the next few years,

but only completed theses were included in this review. In addition, we found two theses published abroad (Professor Dobson and Professor Smith). Even though they both are in the scope of this review, we have decided to include only dissertations from Norwegian institutions in this review because it would be impractical to track down all doctoral theses completed abroad.

Thematically, the doctoral work covered a broad range of topics: portfolio assessment, identity and assessment, mathematics and assessment, assessment in foreign languages and dialogue for learning. The Norwegian search databases we used in our search only uncovered six published doctoral theses. The remaining 23 theses were found through our network. It follows that there is limited value in using relevant research databases to identify Norwegian dissertations. Keywords and abstracts did not necessarily match and some of the dissertations were not available online. However, reviewing published doctoral work on assessment and learning in Norway over the past 20 years, the conclusion has to be that assessment and learning is an active field of doctoral research, which bodes well for the future of this area.

From 1993 until March 2014, 17 doctoral dissertations were published in Norway which included assessment and learning. From 2003 until 2014, a total of ten dissertations were published on the topic of national or international tests alone. In other words, the influence of international tests can also be seen on the dissertation work conducted in Norway. These theses used were based upon data from TIMSS (Naalslund, 2012 – Table 5), TIMSS and PISA (Olsen, 2005 – Table 5), PISA (Turmo, 2003; Hopfenbeck, 2009 – both in Table 5), International Adult Literacy Survey (IALS) (Gabrielsen, 2003 – Table 5) and the ALL study (Lundetræ, 2010 – Table 5). Only one dissertation used data from national tests (Nortvedt, 2010 – Table 5).

Table 2 Number of doctoral theses on assessment in Norway since 1993

Institutions	Assessment and learning	International tests	National tests	Exams	Total
University of Oslo	6	5	1	1	13
University of Bergen	3	1	0	0	4
Arctic University of Norway (Tromsø)	1	0	0	0	1
University of Agder	2	0	0	0	2
Norwegian University for Science and Technology (NTNU)	4	0	0	0	4
University of Stavanger	1	2	0	1	4
Oslo and Akershus University College of Applied Sciences	0	0	1	0	1
Total	**17**	**8**	**2**	**2**	**29**

The dissertations on national and international tests were all published in the past ten years, while the dissertations on assessment and learning had a longer trajectory. The

increase of dissertations on national and international tests can partly be explained by the Ministry of Education policy, which financially supported dissertations on international datasets as a part of the participation of international comparative studies. Both the Unit for Qualitative Analysis in Education (EKVA) at the University of Oslo and the National Reading Centre at the University of Stavanger are in charge of these international comparative studies by doctoral students working with these datasets. It is also a natural consequence of the introduction of national tests in Norway in 2004, as well as the first PISA cycle in 2001. In the 1990s, Norway only participated in TIMSS, and the focus in assessment research was more upon school-based evaluation and less focused upon tests (Haug & Monsen, 2002; Imsen, 2004).

In the next three tables, each dissertation is listed, with the name, title, affiliation and year published, followed by a final section showing abstracts from each dissertation. Not surprisingly most of the dissertations have been published at the University of Oslo (12), which is the largest university in Norway, followed by the second largest university, the University of Bergen, University of Stavanger and Norwegian University for Science and Technology (NTNU) each with four published theses.

Table 3 Doctoral theses on assessment and learning or evaluation

	Name, year and institution	Title
1	Allern, Marit Kristin (2005) University of Tromsø	Individuell eller kollektiv læreprosess? Mappevurdering i praktisk-pedagogisk utdanning (Individual or collective learning processes? Portfolio assessment for the Graduate Diploma in Education)
2	Eggen, Astrid Birgitte (2005) University of Oslo	Alfa and Omega in Student Assessment. Exploring Identities of Secondary School Science Teachers
3	Emstad, Anne Berit (2012) Norwegian University for Science and Technology NTNU	Rektors engasjement i arbeidet med oppfølging av skolevurdering: En kvalitativ kasusstudie av hvordan seks norske barneskoler har brukt skolevurdering i sitt arbeid med forbedring av skolen som læringsarena (Principles engagement in use of evaluation: a qualitative case study of how six Norwegian schools have used school evaluation in their work for improving the school as a learning arena)
4	Engelsen, Knut Steinar (2006) University of Bergen	Gjennom fokustrengsel: lærerutdanningen i møte med IKT og nye vurderingsformer (Teacher education meeting ICT and new ways of assessment)
5	Gamlem, Siv Måseidvåg (2014) University of Stavanger	Feedback as support for learning. A classroom study on how feedback can be perceived practiced and used by pupils and teachers on secondary school
6	Gynnild, Vidar (2001) NTNU	Læringsorientert eller eksamensfokusert?: Nærstudier av pedagogisk utviklingsarbeid i sivilingeniørstudiet (Learning oriented or examoriented? In depth studies of pedagogical development work in Master of Science in Business)

7	Hanken, Ingrid Maria (2007) University of Oslo	Studentevaluering av individuell hovedinstrumentundervisning. En caseundersøkelse av en høyere musikkutdanningsinstitusjon (Student evaluation of individual instrument teaching. A case study of a higher education Institute of Music)
8	Isabwe, Ghislain Maurice Norbert (2013) University of Agder	Enhancing Mathematics Learning through Peer Assessment Using Mobile Tablet Based Solutions
9	Mausethagen, Sølvi (2013) Oslo and Akershus University College of Applied Sciences	Reshaping teacher professionalism. An analysis of how teachers construct and negotiate professionalism under increasing accountability
10	Munkebye, Eli (2012) University of Oslo	Dialog for læring – den utforskende naturfaglige samtalen i uteskole (Dialogue for learning – the explorative natural science dialogue in outdoor education)
11	Prøitz, Tine (2014) University of Oslo	Conceptualisations of learning outcomes in education – an explorative cross-case analysis of policymakers, teachers and scholars
12	Randen, Gunnhild Tveit (2014) University of Bergen	Tilstrekkelige ferdigheter i norsk? Kartlegging av minoritetsspråklige skoleelevers språkferdigheter. (Adequate skills in Norwegian? Mapping of immigrant pupils' language skills)
13	Roald, Knut (2010) University of Bergen	Kvalitetsvurdering som organisasjonslæring mellom skole og skoleeigar (Quality Assessment as oragnisational learning between school and school owner)
14	Sandvik, Lise Vikan (2011) NTNU	Via mål til mening: En studie av skriving og vurderingskultur i grunnskolens tyskundervisning (From goals to meaning: a study of writing and assessment culture in German undergraduate teaching)
15	Sjøbakken, Ola Johan (2012) University of Oslo	Elevsamtalen som jevnlig dialog i et aksjonsforskningsperspektiv (Frequent teacher–pupil dialogue as in an action research perspective)
16	Skedsmo, Guri (2009) University of Oslo	School governing in transition? Perspectives, purposes and perceptions of evaluation policy
17	Wheat, David (2004) University of Bergen	The feedback method. A System Dynamic Approach to teaching macroeconomics
18	Wittek, Anne Line (2007) University of Oslo	Portfolio as an artefact for learning in higher education Structures, cultural practice and trajectories of participation

Table 4 Doctoral theses on exams

	Name, year and institution	**Title**
19	Berge, Kjell Lars (1996) NTNU	Norsksensorenes tekstnormer og doxa. En kultursemiotisk og sositekstologisk analyse. (Textual norms and doxa of Norwegian Examiners. A cultural semiotic and analysis)
20	Skjelten, Sidsel (2013) University of Stavanger	Jakta på kvalitetsforskjellar i elevane sine tekstar. Kva skil gode tekstar frå middels gode? (The search for writing quality differences in pupil texts. What differentiates very good texts from good texts?)

Table 5 Doctoral theses on international and national tests

	Name, year and institution	Title
21	Gabrielsen, Egil (2003) University of Bergen	Lese for livet. Lesekompetansen i den norske voksenbefolkningen sett i lys av visjonen om en enhetsskole. Data fra IALS (International Adult Literacy Survey). (Reading for life. Reading competence in the Norwegian adult population in light of the vision of a unified school. Data from IALS (International Adult Literacy Survey)
22	Hellekjær, Glenn Ole (2005) University of Oslo	The acid test: does upper secondary EFL instruction effectively prepare Norwegian students for the reading of English textbooks at colleges and universities?
23	Hopfenbeck, Therese Nerheim (2009) University of Oslo	Learning about Students' Learning Strategies. An empirical and theoretical investigation of self-regulation and learning strategy questionnaires in PISA
24	Lundetræ, Kjersti (2010) University of Stavanger	16–24-åringers basisferdigheter: en studie av basisferdigheter relatert til selvoppfatning, frafall i videregående opplæring og arbeidsledighet (16–24-year-olds' basic skills: a study of the basic skills related to self-esteem, dropouts in secondary education and unemployment)
25	Naalslund, Margrethe (2012) University of Oslo	Why is algebra so difficult? A study of Norwegian lower secondary students' algebraic proficiency
26	Nortvedt, Guri Anne (2010) University of Oslo	Norwegian Grade 8 students' competence in understanding and solving multistep arithmetic word problems
27	Olsen, Rolf Vegar (2005) University of Oslo	Achievement tests from an item perspective: an exploration of single item data from the PISA and TIMSS studies, and how such data can inform us about students' knowledge and thinking in science
28	Solheim, Oddny Judith (2010) University of Stavanger	Assessment of reading performance, fundamental conditions for measuring reading
29	Turmo, Are (2003) University of Oslo	Naturfagdidaktikk og internasjonale undersøkelser. Store internasjonale studier som ramme for naturfagdidaktisk forskning: En drøfting med eksempler på hvordan data fra PISA 2000 kan belyse sider ved begrepet naturfaglig allmenndannelse (Science education and large-scale international comparative achievement studies: A thesis for the degree of Doctor Scientiarum)

9.5 Classification by theme

The dissertations described in this review cover all school levels from primary to higher education. Some are subject-specific, such as science and mathematics, music, English and German as foreign langugages and international tests, and some describe more general pedagogical processes such as school development and evaluation. The methods also vary from large-scale to small-scale studies, case studies, interviews, classroom observations and document analysis, but as is typical for educational research more generally, the majority are qualitative dissertations. Thematically, the

dissertations were classified into the ten groups listed below. As there are only 29 dissertations, some of the classifications included a single thesis.

9.5.1 School policy and school evaluation

Four of the theses addresses issues regarding school policy and school evaluation and how school leaders and teachers respond to the increased focus upon learning outcomes such as national tests (Emstad, 2012; Mausethagen, 2013; Roald, 2010; Skedsmo, 2009 – all in Table 3). Emstad, Roald and Mausethagen used different qualitative techniques such as interviews and document analysis while Skedsmo analysed data collected from a national survey directed to school principals using Structural Equation Modelling.

9.5.2 Portfolio assessment and writing

A total of four theses investigated this topic. Two of the theses described the use of a portfolio as a tool for learning and assessment and both researchers collected data from higher education institutes (Allern, 2005; Wittek, 2007 – both in Table 3). Two other dissertations looked at writing and assessment, with a particular focus upon quality in texts and teachers' assessment competencies (Berge, 1996; Skjelten, 2013 – both in Table 3).

9.5.3 International comparative studies: PISA, TIMSS, PIRLS, IALS

Three of the studies looked at reading. One dealt with problems regarding validity in relation to reading comprehension. The researcher used a mixed methods approach and analysed the dataset from PIRLS (Solheim, 2010 – Table 5). The second and third studies analysed data from IALS (Gabrielsen, 2003; Lundetræ, 2010 – both in Table 5). Gabrielsen investigated reading comprehension among adults in relation to what is known to be the Unified School Model in Norway while Lundetræ explored whether reading and mathematical skills were related to students dropping out of schools in Norway. Three studies used datasets from PISA; Olsen (2005 – Table 5) compared PISA items in science to TIMSS items, while Turmo (2003 – Table 5) analysed PISA questionnaire data from the first cycle with particular focus upon science, socioeconomic status and learning strategies. Hopfenbeck (2009 – Table 5) validated the student questionnaire in PISA with a mixed methods approach, analysing the learning strategy items used as national options on the PISA test in Norway in 2006. One study (Naalslund, 2012 – Table 5) used TIMSS data to examine students' algebraic proficiency.

9.5.4 National tests

Only one thesis used data from national tests for secondary analysis. Nortvedt (2010 – Table 5) used a sample from mathematic national tests in grade 8 to analyse students' understanding of problem solving, and how it is linked to reading comprehension.

9.5.5 Assessment and ICT

Two theses discussed and analysed assessment and the use of ICT. One analysed new ways of assessment forms and the use of ICT, whilst the other dealt with students developing tablet technology to support peer assessment. The first had a small-scale qualitative design, while the latter used a mixed methods design (respectively Engelsen, 2006; Isabwe, 2013 – both in Table 1).

9.5.6 Feedback method in specific domains

One study concerned a feedback method for improving undergraduate instruction in macroeconomics. The study had a quantitative experimental design (Wheat, 2004 – Table 3). A second study looked at how students are given feedback in music, when they are playing instruments to become professional musicians (Hanken, 2007 – Table 3). The study built upon interviews of nine instrumental teachers and nine students and had a qualitative research design.

9.5.7 Dialogue between teachers and pupils

Two theses related to dialogue between teachers and pupils: one in natural science outdoors education, the other on explorative dialogues and teachers' choice of strategy to support pupils' learning. Both studies were qualitative: one a longitudinal study with an action research approach where a group of teachers were followed over three years, the other analysis of five teachers' dialogue with their pupils (Sjøbakken, 2012; Munkebye, 2012 – both in Table 3).

9.5.8 Language assessment

Two studies addressed the issue of assessment in a foreign language, in this case German and English as second languages. One study used mixed methods, and included interviews, observations and students' texts. In addition a questionnaire was reported (Sandvik, 2011 – Table 3). Randen (2014 – Table 3) investigated immigrant students' language skills in Norway, while Hellekjær (2005 – Table 5) investigated English as a foreign language (EFL) instruction and Norwegian students' reading comprehension in English.

9.5.9 Students' learning orientation and assessment

Gynnild (2001 – Table 3) conducted an action research project where a mid-term exam was implemented to improve students' learning. The study addressed students' learning orientations and study habits in relation to exams and assessments. A mixed methods design was used.

9.5.10 Assessment and identity

One thesis explored the identity formation of secondary school science teachers (Eggen, 2005 – Table 3). The thesis addresses epistemological and ideological dilemmas of student assessment as teacher identity formation. This was a qualitative ethnographic study involving ten teachers from Norway, Sweden and England.

Thesis abstracts

We list the abstracts by author alphabetically, in Norwegian and English where available.

Marit Kristin Allern: Individual or collective learning process? Portfolio assessment for the Graduate Diploma in Education

Abstract
Portfolio assessment is generally understood as a form of assessment replacing traditional exams. This is often the case, but portfolio assessment ought to be a part of an overall pedagogy. If not arrangements called portfolio assessment might be just more of traditional assignments placed in a portfolio.

The dissertation is a study of the implementation of portfolio assessment as a way of working and an assessment form in Graduate Diploma Education Course at the University of Tromsø. It is a qualitative case study. The data set consists of interviews with students and teachers, surveys among students, observation, the students' reflective logs and the introductions to the portfolios. The data was collected in an authentic context from February 2000 till June 2002. The dissertation builds on a sociocultural perspective on learning and learning in a community where collective and individual work is attempted balanced and there is a focus on consequences for teaching and learning activities.

To implement portfolio assessment the way it was attempted here, is a demanding process making high demands to the participants. The resources of portfolio assessment can be found particularly in the relation student-student. The challenge is to trigger off this resource which calls for high extent of cooperation between students,

among students and teacher(s) and among teachers, a collective implementation also supporting the individual's development.

The teacher's task is offering help and support in the zone of proximal development and to contribute to essential scaffolding. If an extended metaphor is used, this means to develop a dialogic interaction. This implies big challenges because it is central that the students eventually are supposed to take over the control over their own learning and development process. Traditional learning culture is hard to change and both teachers and students are characterized by it, it is difficult to depart from habits for all persons involved.

Portfolio assessment is not recommended in every context. To what extent it will be a success depends on several factors and depends for example on a group of teachers wanting to make an effort together and who decide to go for a collective *implementation.*

Astrid Birgitte Eggen: Alfa and Omega in Student Assessment: Exploring Identities of Secondary School Science Teachers

Abstract

The dissertation is aiming at investigating the identity formation of the secondary school science teachers. The core phenomenon is student assessment, and more specific epistemological and ideological assets and dilemmas of student assessment as teacher identity formation. Student assessment has grown in number of purposes and has been extended due to institutionalization as well as multifold educational agendas. However, student assessment seen as the challenges teachers are facing and their reflection about student assessment has had less focus empirically. This dissertation is attempting at capturing some of these challenges by combining two theoretical foci and traditions; science education and pedagogy. The main research question and four sub-questions are:

> *Within an overall sociocultural view on reflective identity formation what are the assessment dilemmas, epistemological and scientific ideological viewpoints that constitute the science teacher's student assessment practices and corresponding reflections?*

A. Research questions concerning the science teacher actions and reflections as studied by participant observation, interviewing and the analytical techniques of Grounded theory: What are the teachers' implicit and explicit epistemological and ideological assessment dilemmas?

B. Research questions regarding the development of reflection: What specific epistemological and ideological dilemmas can be identified as a part of the teacher's identity in the single education situation and during the course of the fieldwork?

C. Research questions pertaining to the varieties of identities within science education as analyzed within Grounded theory: What are the different epistemological and ideological assessment dilemmas that can be identified using constant comparative cross-case analytical methods among the participating teachers?
D. Research questions regarding the application of ethnography and Grounded theory methodologies: What are the methodological considerations concerning the investigation of science teachers' actions and reflections concerning student assessment?

The empirical study is an ethnographic study based on participant observations, formal interviews and discourses with ten teachers in Norway, Sweden and England. Five of these teachers are presented as case-records. They are selected for the purpose of illustrating ideological and epistemological positions and dilemmas regarding student assessment and they hence form a typology labeled Alfa, Gamma, Pi, Sigma and Omega.
Following a Grounded theory methodological approach, the theoretical framing has been developed in co-operation with practicing teachers. One of the teachers, Delta is a sign-poster throughout the theoretical discussions of dilemmas of assessment, epistemological implications for assessment and science ideological positions.
Among the dilemmas that are being theoretically elaborated are:

Formative versus summative assessment.
Ipsative (individual), group or objectives as references for assessment.
Autocracy or autonomy in assessment practices.
Scientific knowledge as universal, defined and given or knowledge as continuously developed in communities
Learning as individual processes or learning as collective and relational processes.
The student as a client of assessment or the student as active participant in forming the assessment references.
Science content as given according to the academic discipline structure or science content as based on societal and student life challenges.

Alfa is ideologically and epistemologically non-dualistic. His essentialistic and behaviorist assessment practice that contains summative purposes does not relate to dilemmatic reflection. Gamma is the manager of assessment and curricula guidelines. His reflections are based on mandated documents. Pi's identity is ideologically and epistemologically dualistic. His concern is to build on individual cognition for learning while his assessment procedures are entirely based on summative testing traditions. He does not acknowledge this dilemma, but emphasize organizational solutions like ability grouping and additional grading scales to capture effort and conceptual learning. Sigma is ideological and epistemological diverse. She recognizes dilemmas of student assessment and sees these as sources for reflection and professional development. Finally, Omega is the progressivist teacher with sociocultural perspectives on learning as well as entirely concerned with formative purposes of assessment.

The dissertation is presenting the complexity of student assessment, and how the teachers are positioning themselves according to different educational contexts. They construct and reconstruct different identities ideologically and epistemologically. These are important dimension in order to develop knowledge and reflection about student assessment within the present multifold objectives for education. Such increased student assessment awareness is alpha and omega for the teaching of science as a knowledge area.

Anne Berit Emstad: Principles engagement in use of evaluation: a qualitative case study of how six Norwegian schools has used school evaluation in their work for improving the school as a learning arena

Sammendrag
Avhandlingen tar for seg rektor sitt engasjement i bruk og oppfølging av skolevurdering. Hensikten med studien er å utvikle kunnskap om de prosessene som har foregått i etterkant av vurderingsarbeidet. Videre er hensikten å få en dypere forståelse av hvordan skoler bruker skolevurdering i sitt arbeid med forbedring av skolen som læringsarena, og hvordan rektor engasjerer seg i dette arbeidet. Avhandlingen er bygd opp rundt fire artikler som presenterer og diskuterer funnene i studien. Studiens overordnede problemstilling er:
Hvordan brukes skolevurdering i skolen som grunnlag for videre arbeid, og hvordan engasjerer rektor seg i dette arbeidet?

Studien er en kvalitativ kasusstudie med bruk av fenomenologiske intervjuer med rektor, lærere og elever ved seks barneskoler som har gjennomført en skolevurdering. Jeg spurte hvordan lærerne hadde erfart rektors engasjement i arbeidet med bruk og oppfølging av skolevurderingsresultatene, og intervjuet også rektorene om hvordan de erfarte sin egen rolle i arbeidet med bruk av skolevurdering. I tillegg består det empiriske materialet av observasjoner av møter og dokumentanalyser. Fire av skolene gjennomførte en ekstern skolevurdering, og to av skolene gjennomførte en intern skolevurdering. Vurderingsmodellene som ble benyttet av skolene, var ment å imøtekomme lov og forskrift for grunnskolen og den videregående opplæring, om ansvarsomfang og skolebasert vurdering.

Studiens overordnede problemstilling blir i fire artikler belyst fra ulike teoretiske perspektiver og review av ulike forskningsfelt som har hatt relevans for studien. I artikkel 1 ser jeg på ulike typer bruk av skolevurdering og ulike faktorer som påvirker bruk. Artikkel 2 tar utgangpunkt i ledelsesteori utviklet av Erik Johnsen (2002) og diskuterer ledelse av ledelsesprosesser. Teori om organisasjonslæring benyttes for å belyse problemstillingen i artikkel 3, mens teori og review av forskning på accountability danner rammen for artikkel 4. Diskusjonen er foretatt på grunnlag av funn presentert i de fire artiklene og er rammet inn av Dewey sine teorier om læring og utvikling samt empirisk generert teori om profesjonelle læringsfellesskap og hensikten

med vurdering. Skolevurdering av, for og som forbedring av skolen som læringsarena danner strukturen i diskusjonskapitlet.

Studien viser ulike typer bruk av skolevurdering. Tre av skolene bruker resultatene av skolevurderingen som grunnlag for å ta en beslutning om videre arbeid. En skole bruker vurderingen som en læringsprosess, mens to skoler ikke følger opp resultatene i sitt videre arbeid. Et av hovedfunnene i studien er at rektors engasjement har betydning for bruk av skolevurdering. Rektorenes prioritering, tanker og ønsker om skolens behov for videre arbeid, tilrettelegging og deltakelse i oppfølging av prosessene har hatt betydning for på om og på hvilken måte skolevurderingen brukes.

Et annet funn i studien er at lærerne i liten grad deltar i en kunnskapsbasert refleksjon rundt resultatene av skolevurderingene, og det brukes liten tid på dette før skolene bestemmer seg for hvordan de skal bruke funn i skolevurderingen. «Activity traps» (Earl, Katz, & Ben Jafaar, 2009) er et begrep som innebærer sterkt fokusert handling, der det brukes liten tid på å finne frem til områder av praksis det ønskes å fokusere på. Et viktig element i profesjonelle læringsfellesskaper er evnen til kritisk tenkning og refleksjon. Kritisk tenkning og refleksjon kan bidra til økt profesjonalitet (Dale, 1998). Refleksjonen rundt funn beskrives av forskningsdeltakerne i høy grad som en diskusjon om hva som passer, ikke passer, hva de kan tenke seg å jobbe med, og hva de ikke kan tenke seg å jobbe videre med. Ved å bringe inn ny kunnskap i refleksjonen kunne skolene i større grad utfordret sine erfaringer og handlingsteorier. Dette kunne gitt skolene et bredere grunnlag til å foreta en beslutning om hvilken del av sin praksis de har behov for å følge opp og forbedre.

Abstract
This thesis looks into the engagement of the principals in the use of school evaluation. The purpose of the study is to increase the knowledge of the processes that take place after the schools' have completed the evaluation and to establish a more thorough understanding of how the schools use the evaluations to improve their learning environments. A particular focus is given the role and engagement of the school principals in this work. The thesis is founded on four articles that present and discuss the findings of this study. The overall problem addressed in this study is:
How is school evaluation used as a foundation for further work in the schools and how do the principals engage themselves in this work?

The study is a qualitative case-study using phenomenological interviews with principals, teachers and pupils at six primary schools that had implemented and completed one school evaluation. I asked the teachers how they had experienced the principal's engagement in the use of school evaluation and the follow-up work with the results of the evaluation. I also asked the principals how they had experienced their own role in this work. In addition to the interviews, the empirical material consists of observations of meetings and document analyses. Four of the schools had completed an

external school evaluation and two of the schools had completed an internal school evaluation. The evaluation model used by the schools was expected to meet the Norwegian Education Act and its Regulation, about accountability and school evaluation.

The overall objectives of this study are discussed from different theoretical perspectives and reviews of various fields of research with relevance for the study. In the 1st article, I look at different ways of using the school evaluation and various factors that may influence the use. The 2nd article is based on management theory developed by Erik Johnsen (2002) and discusses the management of the management processes. Theory on organizational learning is used as a theoretical framework of the third article, whereas theory and reviews of research on accountability builds the frame of the 4th article. The discussion is based on the findings presented in the four articles and is framed by Dewey's theories on learning and development as well as empirically generated theory on professional learning community and the purpose of evaluation. The use of school evaluation of, for and as school improvement of the school as a learning environment forms the structure of the discussion.

This study shows different ways of using the school evaluation. Three of the schools use the school evaluation as a basis for further decisions on future work. One school uses the evaluation as a learning process, whereas two schools do not take the evaluation results into consideration in their further work. One of the main findings in this study is that the principals' engagement is of significant matter as to how the school evaluation is used. The principals' thoughts, wishes and rank of priorities when it comes to the schools' needs for further adjustments and participation in the follow-up of the processes, does influence if and in which way the school evaluation is used.

Another finding in this study is that the teachers' participation in reflecting on the school evaluation results is negligible, and that only a minimum of time is spent on this when the schools decide how to use the evaluation results. «Activity traps» (Earl, Katz, & Ben Jafaar, 2009) is a concept that involves a strongly focused action where little time is used to find desired areas of practice. An important element in professional learning community is the ability to think and reflect critically. Critical thinking and reflection may raise the level of professionalism (Dale, 1998). The research subjects of this study describe the reflections on evaluation findings as a discussion about what fits in/ what does not fit in, and in which areas do they/ do they not have an interest in working with improvement. In bringing new knowledge into this reflection, the schools might challenge their experiences and their theories of action. This could give the schools a better foundation in their decision-making as to which part of their practice they need to follow up and improve.

Knut Steinar Engelsen: Teacher education meeting ICT and new ways of assessment

Sammendrag
Som deltakar i det nasjonale endringsprosjektet PLUTO (2000–2003) gjekk lærarutdanninga ved Høgskolen Stord/Haugesund gjennom omfattande endringar. Desse endringane var først og fremst knytte til meir bruk av IKT (Informasjons og kommunikasjonsteknologi), meir studentaktive læringsmåtar, meir samgang mellom praksis og teori, og meir bruk av mappebaserte lærings- og vurderingsformer. Engelsen sin studie er ein analyse av dei erfaringane som vart hausta gjennom denne endringsprosessen. Analysen fokuserer spesielt på koplinga mellom IKT og mappebaserte lærings- og vurderingsformer.

Med støtte i eigen empiri og nyare forsking på området, konkluderer Engelsen sin studie med at dersom IKT skal spele ei viktig rolle i eit læringsmiljø, er det viktig først å skape legitimitet for teknologien gjennom den ordinære faglege aktiviteten, og faga sine norm- og verdisett. Slik legitimitet synes først å bli skapt når aktørane, på ein ekte måte, opplever IKT som nyttig i lærings- og undervisningsarbeidet. Ein naturleg konsekvens av dette er at opplæring på bruk av IKT-reiskapar må ha eit langsiktig perspektiv, og knytast nært opp til dei faga og den læringsaktiviteten reiskapane skal nyttast i. I dette ligg også at isolerte kurs i bruk av IKT-reiskapar, frigjort frå den ordinære faglege konteksten, synes å ha liten effekt.

Engelsen konkluderer vidare i analysen sin med at det synes å vere svært viktig at dei vurderingsmåtane som vert nytta, spelar saman med og bygger opp under dei pedagogiske endringsprosessane. Erfaringar frå lærarutdanninga ved Høgskolen Stord/Haugesund tyder på at mappebaserte lærings- og vurderingsformer, under visse vilkår, kan spele godt saman med studentaktive læringsformer og intensiv bruk av IKT.

Egil Gabrielsen: Reading for life. Reading competence in the Norwegian adult population in light of the vision of a unified school. Data from IALS (International Adult Literacy Survey)

Sammendrag
Lesekompetanse har vært i søkelyset i de senere årene, og det er gjennomført mange både nasjonale og internasjonale undersøkelser av barn, unge og voksnes leseferdigheter. Denne avhandlingen har sin bakgrunn i den internasjonale kartleggingen av leseferdighet i aldersgruppen 16-65 år, IALS ("International Adult Literacy Survey") som ble gjennomført i perioden 1994-98. I alt 22 land var med i undersøkelsen som ble gjennomført som hjemmebesøk hos et representativt utvalg av personer fra den aktuelle aldersgruppen. Over 5000 voksne i Norge deltok i undersøkelsen gjennom å la seg intervjue og ved å arbeide med et utvalg av leseoppgaver.

Avhandlingen ser nærmere på i hvilken grad det finnes sammenhenger mellom lesekompetansen hos ulike aldergrupper av voksne og det norske samfunnets vektlegging av å skape et likeverdig skoletilbud for alle; uavhengig av bosted, kjønn, målform og sosial bakgrunn. I avhandlingen er dette belyst både med bruk av "nasjonale briller" og gjennom en internasjonal sammenlikning av leseferdighet hos voksne.

Gamlem, Siv Måseidvåg: Feedback as support for learning. A classroom study on how feedback can be perceived practiced and used by pupils and teachers on secondary school

The purpose of this research study is to increase knowledge about how feedback to support learning is practiced and perceived by students and teachers in lower secondary school. Through classroom research the study is rooted in a context where feedback will be given, received, sought, used or discarded. Observations and a student- and teacher perspective on feedback practices are presented. An intervention on teachers' learning in the development of feedback practice is also conducted. The overarching research question is: What characterizes feedback practices in lower secondary school, - and how is feedback to support learning perceived and used by students and teachers?

The study is rooted in social constructivism, where individuals are seen as active and responsible for their development, and knowledge is perceived as a construction of understanding and meaning between people. The concept feedback builds on the theory of Assessment for Learning (AfL).

Data are collected during the school year 2009–2010 in four lower secondary schools. Data material consists of video observations, interviews and an intervention where workshops, observations and video-stimulated reflective dialogues are conducted. The research study has an exploratory mixed methods design with an emphasis on qualitative method, where both quantitative and qualitative analyses are used. The approach is designed to be able to gain knowledge of how feedback is presented and perceived in learning activities by teachers and students, and how teachers might change perceptions and practice of feedback over a school semester.

Results from video observations show that the quality of oral feedback interactions in classroom between teacher and student is on average low when it comes to professional instructional support and contributions to strengthen students' meta-learning, but on average high when it comes to emotional support.

Students (age 13–15) emphasize three dimensions that are important for feedback to support learning. They talk about the feedback valence (positive-negative feedback), relationships and honesty when giving/receiving feedback, and different types of

feedback. Students' perceptions are that the feedback is negative when the student (i) is told he could have done a better job even when he thinks he has done his best, (ii) is told he must work more and better, but not how, (iii) receives feedback after the work is "finished" and cannot use the information to improve the product or performance, and (iv) receives praise directed at the person, but not information that can support the learning. Students describe feedback as positive and supportive for learning when students get (i) recognition of their work, performance or effort, (ii) specific information on how they can improve the work, (iii) timely feedback, and (iv) the opportunity to follow up and use the information to improve their work. Furthermore, students emphasize that relationships between students will influence the extent to which students are honest in their feedback to their classmates. Peer feedback focuses on positive elements when it concerns students one has a good relationship with, and vice versa. If there are no objective criteria for students to use when assessing peer work (or their own), they end up with personal "likes and dislikes", creating a dilemma for the students. A typology of various feedback types is presented based on student perceptions of classroom feedback to support learning.

A result from the intervention study demonstrates that understanding of feedback that supports learning is not common to teachers at the same school. Teachers have different understandings of feedback that supports learning and practice feedback, individually based on their own understanding. Support for the re-conceptualization of development seems to be important. The intervention study shows that teachers' learning trajectories are individual with regard to feedback based on the individual teacher perception about the concept of feedback and assessment for learning (AfL), and their established practice. Teachers with an established understanding that is consistent with theories/empirical studies presented for feedback to support learning (AfL), are more likely to change, develop and improve in the same direction.

The study contributes to increased understanding of how feedback is perceived and used as a tool for learning, where strengths and weaknesses are revealed. The results point to a need for increasing teachers' knowledge about feedback as a support for learning, and the findings should have consequences for the ways policy makers, teachers and school owners work with feedback as a support for students' learning in scientifically based development projects. Since the results reveal that feedback interaction based on professional instructional support is on average of low quality, this is an indication that teachers need external support in order to further develop their understanding and practices for feedback interactions that support learning. The challenge is to increase teachers' understanding of support for learning through feedback as interactive activity between the agents in the classroom (teacher-student[s], student-student), where there will be stronger emphasis on instructional support through feedback based on quality, negotiation, shared understanding of goals and success criteria, and interaction.

Vidar Gynnild: Geared towards learning or exams? Cases studies of educational development in engineering education

Abstract
Method: This is a mixed methods study based on quantitative and qualitative data.
This study was conducted as an action research project within marine engineering at Norwegian University of Science and Technology (NTNU) in Trondheim in the late 1990s. The investigation started out by examining the effects of the ongoing student evaluation of teaching in two courses. Not surprisingly, the application of data collected yielded only minor changes in teaching with no evidence of improved student learning as a result.
This initial investigation served as a backdrop for implementing a mid-term exam accounting for 30% towards the final grade. Also other measures to improve learning were implemented, such as the use of laboratory exercises and computer visualizations.

The study provided evidence of the significance of exams and assessments in general for students' approaches to learning. Some students were clearly driven by a desire for deep learning whereas others were clearly driven by what was required in the exams. While the intention of introducing a mid-term exam was to motivate for enhanced efforts earlier in the semester and thereby improve learning, realities turned out not to as straightforward.

While efforts were generally enhanced prior to the mid-term exam, students still practiced cramming before the exams, and there is little if any evidence of improved learning caused by the mid-term exam. Our reading is that our intervention ran counter to pronounced features of the learning culture, and that there is a need for more comprehensive changes in the overall design of the study program to significantly change students' approaches to learning.

Ingrid Maria Hanken: Student evaluation of individual instrument teaching. A case study of a higher education Institute of Music

Sammendrag
Norske musikkutdanningsinstitusjoner befinner seg i spenningsfeltet mellom manghundreårige mesterlæretradisjoner og dagens krav til studiepoengproduksjon, "accountability" og kvalitetssikringssystemer. I dette spenningsfeltet skal den obligatoriske studentevalueringen av undervisning i ulike musikkfag finne sin plass og funksjon.

Dette høgskolepedagogiske forskningsprosjektet har som bakgrunn de utfordringer som oppsto når systematisk studentevaluering ble innført i den individuelle hovedinstrumentundervisningen ved Norges musikkhøgskole. For å kartlegge og belyse disse utfordringene ble det foretatt intervjuer med hovedinstrumentlærere og deres

studenter med utgangspunkt i følgende problemstilling: Hvordan oppfatter, praktiserer og erfarer hovedinstrumentlærere og deres studenter studentevaluering av individuell hovedinstrumentundervisning?

Viktige funn i undersøkelsen er at to konstituerende faktorer ved hovedinstrumentundervisningen kan virke hemmende for studentevalueringen: For det første, studentenes behov for å ha en nærmest uforbeholden tillitt til hovedinstrumentlærerens autoritet: Kravet om å innta en distansert og vurderende holdning til undervisningen kan oppleves som uforenlig med en slik tillit. For det andre, behovet for å ha en velfungerende og uanstrengt relasjon til hovedinstrumentlæreren: Redselen for å såre eller irritere læreren gjennom evalueringene kan virke hemmende på studentenes villighet til å gjennomføre evalueringer. Videre tyder resultatene på at de metodene som velges for å gjennomføre evalueringene har stor betydning for om studentevalueringen blir gjennomført i denne formen for undervisning.

Prosjektet har ført til endringer i prosedyrene og metodene for studentevaluering ved Norges musikkhøgskole. Det har også resultert i at det er utarbeidet en håndbok for studenter og lærere for å imøtekomme de spesielle utfordringer som begge parter står overfor når den individuelle hovedinstrumentundervisningen skal evalueres.

Glenn Ole Hellekjær: The acid test: does upper secondary EFL instruction effectively prepare Norwegian students for the reading of English textbooks at colleges and universities?

Abstract

The present quantitative, descriptive and exploratory study investigates whether, and to what extent, Norwegian upper secondary EFL instruction prepares for the reading of English texts and textbooks in higher education. It uses questionnaires, and a combination of self-assessment items and an academic English reading test (IELTS) to measure English reading proficiency. The samples comprise student respondents from the university and college level as well as senior upper secondary level students from the General Studies branch. Test scores of the senior upper secondary school respondents from the General Studies branch revealed that two thirds would not manage the level required for admission to universities in English speaking countries. Likewise, test and self-assessment scores of university level respondents indicated that reading problems persisted in higher education, with between 30 and 40 percent of the respondents experiencing difficulties.

A closer analysis revealed that the difficulties experienced by many respondents were due to poor language proficiency, exacerbated by a counterproductive tendency towards careful reading with excessive focus on ascertaining the meaning of unknown words. The respondents who indulged in the extracurricular reading of English or had

had Content and Language Integrated Learning courses were among those with the highest scores. Rather unexpectedly, completing the upper secondary level Advanced English Course did not give an advantage. Nor did study experience. Though the findings in this descriptive and exploratory study need to be confirmed in follow-up studies, they clearly indicate the urgent need for changes in the syllabi and teaching of Norwegian EFL instruction

Therese Nerheim Hopfenbeck: Learning about Students' Learning Strategies: An empirical and theoretical investigation of self-regulation and learning strategy questionnaires in PISA

Abstract
The dissertation discusses problems related to the measurement of learning strategies as conducted in Programme for International Student Assessment (PISA), and differences between measuring students' learning strategy use globally or subject specific with questionnaires.

One hypothesis tested is that students that score poorly on the PISA test are less likely to understand the items in the student questionnaire regarding learning strategies.

Data from the piloting of the PISA test in Norway in 2005 (N=1200) and the PISA main survey in 2006 (N=4700) is analysed using factor analysis, correlation analysis and regression analysis. In addition, 22 students who sat for the PISA test in 2006 are interviewed. These cognitive interviews focused upon students understanding of the learning strategy items to control whether they had checked off for valid responses.

The main findings from the survey analysis are that the correlation between learning strategies and scores in reading, Mathematics and Science and confirm previous results from PISA 2000 in Norway. It is not found particular differences between measuring strategies globally vs. subject specific. Neither is if found significantly gender differences in strategy use, but among minority students, girls report to use learning strategies more than minority boys.

The hypothesis of poor performing students and response patterns are confirmed. The lower the students score, the more likely it is that they check off for the different responses randomly, or use the same response category for each question. The interviews also revealed that the lower scores students had on the PISA test, the more challenging they found it to understand the learning strategy items. Both the qualitative and quantitative results indicate that students have not reported valid responses to the learning strategy questionnaires.
The results are of importance for development and interpretation of data from questionnaires.

Sammendrag
Avhandlingen tar opp problemstillinger knyttet til målinger av læringsstrategier ved bruk av spørreskjemaer i PISA, og forskjeller mellom å måle strategier generelt eller i en spesifikk kontekst. Hypoteser som behandles, er i hvilken grad elever som skårer svakere på de faglige PISA oppgavene, er i stand til å forstå innholdet i spørsmålene om læringsstrategier. Det er brukt data fra piloteringen av PISA i 2005 (N=1200) og PISA-undersøkelsen i 2006 (N=4700) og gjennomført faktoranalyser, korrelasjonsanalyser, og regresjonsanalyser. I tillegg er det gjennomført kognitive intervjuer av 22 elever som deltok i PISA. Intervjuene belyste elevenes forståelse av de spørsmålene om bruk av læringsstrategier som de hadde svarte på i spørreundersøkelsen.

De viktigste konklusjonene i forskningsresultatene er at læringsstrategier og faglig skår, samt sammenhengen mellom disse, i PISA 2006 viser til dels samme resultater som i 2000. Det er med andre ord ikke påvist store forskjeller mellom det å måle strategier generelt som i 2000, og det å måle strategier spesielt, slik som i 2006. Det er heller ikke funnet store kjønnsforskjeller blant elever generelt, men blant minoritetselever oppgir jentene at de bruker læringsstrategier i større grad enn guttene.

Det er derimot vist at jo svakere elevene skårer i PISA, jo høyere sannsynlighet har de for å velge samme responskategori på samtlige spørsmål i undersøkelsen om læringsstrategier. Intervjuene viser at elevene med svakest skår i PISA også har størst problemer med å forklare hva spørsmålene betyr. Begge disse forholdene indikerer at de svakeste elevene ikke har gitt valide svar om egen strategibruk.

Forskningsresultatene er relevante i forhold til utvikling av og tolkning av data fra spørreskjemaer i utdanningsforskning, særlig i forhold til elever på ungdomstrinnet.
Arbeidet er utført ved Institutt for lærerutdanning og skoleutvikling ved Universitetet i Oslo. Avhandlingen er skrevet innenfor fagområdet pedagogikk, som monografi.

Ghislain Maurice Norbert Isabwe: Enhancing Mathematics Learning through Peer Assessment Using Mobile Tablet Based Solutions

Abstract
Higher education is facing unprecedented challenges with an increasing demand for high quality education, driven by tougher global competition. Student numbers are fast growing at most universities, whereas the sources of funding are not proportionally increasing. Subsequently, the teaching staff's workload gets higher and higher hence putting the quality standards at risk. As class sizes increase, it becomes more difficult to learn in a highly teacher-controlled environment, since the teacher cannot sufficiently address individual student's needs. Therefore, a teacher should be conceptually seen as a facilitator for students, who provides them with guidance and opportunities to explore and make sense of their subjects of study. Sustainable quality education requires novel approaches to teaching and learning, to provide the best education with

a minimum amount of resources. For instance, students should be encouraged to be more active in their learning rather than being passive receivers of the instruction. This study calls for a fundamental shift from instructionism (a teacher focussed educational practice) towards constructivism (a student focussed educational practice); keeping in mind that a combination of both practices may be needed in certain cases. In addition to adopting the appropriate educational praxis, innovations in educational technology can further enhance the learning experience. Mobile media tablets are gaining popularity with university students as technology matures. Besides communication and digital media consumption as their primary functions, the latest mobile media tablets can also be used for data production and processing in teaching and learning contexts. This work revisits the practice of peer assessment as a means of formative assessment. Based on user centred design principles, engineering students at the University of Agder (Norway) and Kigali Institute of Science and Technology (Rwanda) were involved in developing a media tablet technology supported peer assessment system. The students' role in system development is reported as well as their active learning through peer assessment of mathematics assignments. Results of the experimental study generally showed improvements in the technical usability of the system throughout the development cycle. Analysis of the pedagogical usability criteria suggests that there are possible learning gains of using such a system. The findings indicate that peer feedback has a potential to improve students' learning achievement and that media tablets hold a promising solution in learning mathematics and related subjects. Furthermore, the challenges of implementing effective peer assessment systems supported by new information and communication technology are discussed.

Kjersti Lundetræ: 16–24 year olds' basic skills: a study of the basic skills related to self-esteem, dropout in secondary education and unemployment

Sammendrag

Et samfunn i rask endring har resultert i et forsterket fokus på betydningen av grunnleggende ferdigheter og livslang læring. Det er en sterk sammenheng mellom ungdoms utdanning og kompetanse og deres tilgang til arbeidsmarkedet. Derfor blir bortvalg av og frafall fra videregående skole ansett som en stor utfordring i norsk så vel som i internasjonal kontekst.

Med utgangspunkt i data fra den internasjonale undersøkelsen Adult Literacy and Life Skills Survey bidrar avhandlingen særlig til å utvikle ny kunnskap om sammenhenger mellom basisferdigheter i lesing og regning og frafall i videregående opplæring og arbeidsledighet blant 16–24-åringer.

Resultatene bekrefter blant annet at unge som har mødre med høyt utdanningsnivå har større sannsynlighet for å gå videre i utdanningssystemet enn andre unge. Når man derimot kontrollerer de unges basisferdigheter, viser resultatene at foreldres utdanningsnivå ikke har betydning for fullføring av videregående opplæring i Norge, i

motsetning til i USA. Resultatene viser også at unge med lave basisferdigheter er langt mer utsatt for arbeidsledighet, selv når de har fullført videregående opplæring.

Abstract
The aim with this PhD-project was to provide new knowledge about 16–24-years olds' functional literacy and numeracy skills, with a basis in data from Adult Literacy and Life Skills Survey (ALL). ALL is a comparative, international cross-sectional survey of 16–65-years olds' functional literacy and numeracy skills. The survey was managed by Statistics Canada in cooperation with OECD. In 2003 data was gathered from Bermuda, Canada, Italy, Norway, Switzerland and USA, and data from the five last-mentioned countries was used in different papers in this doctoral thesis. In 2006/2007 ALL 2 was carried out, and data was gathered from Australia, Korea, Netherlands, New Zealand and Hungary. Data from these countries will be presented during the autumn 2010.

The aim of this project was to explore certain problems related to education and work in the different countries by using ALL-data. To decide which valid comparisons that could be made based on ALL, a critical discussion on ALL as a measuring instrument was required. Therefore, paper I is a theoretical paper discussing which valid comparisons that can be made based on ALL. The discussion in paper I was limited to measuring reading skills, translation and adaption, culture and context and motivation. ALL was found to be better suited to compare correlations within countries than reading skills across countries, because of some uncertainties connected with the measuring instrument, despite the survey's high methodological quality. These uncertainties can be avoided by only comparing correlations within countries between countries, as exemplified by paper II to IV.

In paper II it was examined whether there were differences between 16–24-years old girls' and boys' mathematical self-concept when it was controlled for numeracy skills. The results showed that boys had slightly better mathematical self-concept than girls within all the ALL-countries. The gender-differences were smaller, but still significant when controlled for numeracy skills. Despite significant gender differences, the correlations were relatively weak, and the gender differences were assumed to be of little practical meaning. The gender differences were largest in Switzerland and Norway, which is in line with previous research (OECD, 2004).

In paper III it was examined whether parents' educational level had impact on 16–24-years olds' dropout from upper secondary school and training in Norway and USA, when controlled for the youths' basic skills. It was found that parents' and especially mother's educational level impact dropout or early school leaving in upper secondary school and training. When it was controlled for the youths' basic skills, mother's education was no longer significant in Norway, while it had a relatively large impact on dropout in USA.

In paper IV it was examined whether basic skills in terms of literacy and numeracy skills predicted youth unemployment (16–24-years) in Canada, Italy, Norway and USA. The analyses showed that youths with basic skills on the lowest level, were far more exposed to unemployment. In Canada and USA low basic skills even increased the odds of being unemployed when controlled for accomplished upper secondary school and training. As for Norway, when controlling for educational level, basic skills were nearly significant (p = .06), showing an odds ratio of 2.71.

Paper I sheds light on and discuss different aspects of comparisons of skills in international reading surveys. Paper II–IV are examples of comparisons that can be done with data from such surveys, to avoid uncertainty related to e.g. problems related to translation, familiarity, type of item or different motivation for the test. The last three papers also contribute with more knowledge in the field of special education, both when it comes to the importance of adequate functional literacy- and numeracy skills to prevent youth unemployment and dropout or early school leaving and gender differences in mathematical self-concept.

Sølvi Mausethagen: Reshaping teacher professionalism. An analysis of how teachers construct and negotiate professionalism under increasing accountability

Sammendrag

I denne avhandlingen studerer jeg hvordan lærerprofesjonen i Norge fyller 'lærerprofesjonalitet' med innhold. Studien tar utgangspunkt i det økte politiske fokuset på elevresultater og ansvarliggjøring av læreren (accountability), og undersøker om og i hvilken grad slike endringer påvirker oppfatningen av profesjonalitet. Det er lite kunnskap om hvordan lærere i Norge forholder seg til mer ekstern kontroll av arbeidet sitt. Tidligere internasjonale studier av endringer i lærerprofesjonalitet de siste to tiårene har i stor grad basert seg på dokumentanalyse eller intervju. I denne studien består det empiriske datamaterialet av tre stortingsmeldinger, politiske dokumenter fra Utdanningsforbundet, deltakende observasjon av lærermøter, fokusgruppeintervjuer og individuelle intervjuer med lærere, samt 28 fagfellevurderte artikler. I avhandlingen anvender jeg teoretiske perspektiver på profesjon og profesjonalitet, og hvordan utdanningspolitikk utføres lokalt. Det brukes en diskurs-inspirert tilnærming for å undersøke i hvilken grad og på hvilke måter den utdanningspolitiske diskursen står i et spenningsforhold til lærernes diskurs. Sammen muliggjør disse teoriperspektivene diskusjoner rundt forholdet mellom konstruksjoner av profesjonalitet gjort av ulike aktører, og hvordan språket brukes til å skape legitimitet og relevans for lærerarbeidet.

I den første artikkelen undersøker jeg hvordan myndighetene og Utdanningsforbundet har konstruert lærerprofesjonalitet i løpet av det siste tiåret spesielt. Begge aktørene er i økende grad opptatt av profesjonalitet, men har ulike oppfatninger av begrepet. Myndighetene legger vekt på at lærerne skal holdes ansvarlige for elevenes læring, ha en forskningsbasert praksis og være faglig kompetente og oppdaterte. Utdanningsforbundet er på sin side opptatt av yrkesetikk, av forskningsinformert

praksis, samt av at læreren aktivt tar ansvar for kvaliteten i skolen, og deres oppfatninger er dermed tettere knyttet til de klassiske profesjonsidealene. Utdanningsforbundet gjør hovedsakelig motstand mot en ekstern kontroll av lærernes arbeid, men fremstår som mer proaktive enn tidligere siden de vektlegger forskning for å styrke læreres profesjonalitet og tillit.

I den andre artikkelen undersøker jeg hvordan grupper av lærere lokalt definerer det å være ansvarlig, hva de er ansvarlige for og til hvem. Gjennom å bruke begrepene intern og ekstern accountability som sensitiverende begreper, forsøker jeg å 'åpne opp' ansvarsbegrepet ved å studere hvordan lærerne definerer seg selv og hverandre som en ansvarlig lærer. Å være ansvarlig for elevenes læring, overfor læreplanen og andre forskrifter, samt overfor rektorer og foreldre, blir fremhevet som viktig, særlig av yngre lærere. Erfarne lærere er mer opptatt av å være ansvarlige overfor de bredere utdanningsmålene, samt egen erfaring og kunnskap, hvilket igjen benyttes til å delegitimere ekstern accountability. I dette spenningsforholdet mellom det som kan beskrives som og oversettes til 'å være ansvarlig' og 'ansvarliggjøring', har det imidlertid utviklet seg en alternativ legitimeringsdiskurs rundt det å være oppdatert på og bruke forskning i lærerarbeidet.

I den tredje artikkelen utdyper jeg hva som skjer når nasjonale prøver gjennomføres lokalt, og hvordan lærere på lærermøter diskuterer sin praksis rundt disse prøvene. Spenningene som skapes i interaksjonen mellom lærere handler om hva som blir sett på som interne (lærernes daglige arbeid) og eksterne (praksiser plassert utenfor hovedrammen av undervisning) forhold. Det er særlig fire forhold som blir satt 'på spill' for lærere med nasjonale prøver: profesjonskunnskap, læreplanen, formative aspekter ved undervisningen og lojaliteten til elevene. Disse aspektene anses i hovedsak som interne forhold for lærere som deltok i studien. Selv om nasjonale prøver ser ut til og stort sett betraktes som eksterne elementer i arbeidet, involverer lærerne seg i såkalt grensearbeid for å markere hva som er viktig ved egen yrkesutøvelse. Samtidig må lærerne skape relevans ved og legitimere det å øve til prøvene, men dette gjøres gjennom relasjonelle aspekter heller enn å legge vekt på elevenes resultater.

I den fjerde artikkelen tar jeg et internasjonalt perspektiv ved å se på hva eksisterende forskning sier om mulige endringer i lærernes relasjoner til elever og kolleger når ekstern kontroll øker, og spesielt ser jeg på den økende bruken av tester og resultater. Denne studien gir kunnskap om hva som kan være sosiale effekter av standardisert testing slik dette gjennomføres i såkalte 'high-stakes' kontekster. Det vektlegges ofte at mer testing og den betydningen testresultatene får fører til at det rettes mindre oppmerksomhet mot omsorgen til elevene og de relasjonelle aspektene ved arbeidet som lærer. Den samme vektleggingen av positive sosiale relasjoner kan dessuten føre til en motstand mot testene. Relasjoner til kolleger blir også berørt av testing, men både i positiv og negativ retning. Dette peker på betydningen av den organisatoriske

konteksten for lærerens arbeid, og dermed for hvordan man forholder seg til resultatfokus og ekstern kontroll på den enkelte skole.

Funnene i avhandlingen bidrar til å dokumentere endringer i diskurser om lærerprofesjonalitet blant politikere og lærerorganisasjonen, og jeg viser hvordan lærerprofesjonen i Norge synes å ha blitt mer proaktiv i å skape legitimitet for arbeidet sitt. Både Utdanningsforbundet og lærere lokalt gjør i relativt stor grad motstand mot en ekstern kontroll av arbeidet. Denne motstanden formuleres imidlertid kraftigere av Utdanningsforbundet, mens den er mer subtil og også varierende blant lærerne lokalt. For det første synes yngre lærere å være mer positive og balanserte til nye forventninger til lærerrollen. For det andre har en alternativ legitimeringsdiskurs utviklet seg, en diskurs som legger mer vekt på det som kan beskrives som forskningsinformert praksis. Mens accountability hovedsakelig plasseres utenfor lærernes verdisystemer, plasseres (ny) forskning i større grad innenfor. Avhandlingen viser hvordan profesjonen på ulike måter utfører såkalt diskursivt legitimerings- og grensearbeid. Derfor kan et svar på spørsmålet om 'ansvarliggjøringspolitikk' omformer deler av det performative aspektet av læreryrket være ja, delvis. På den ene siden kan det se ut som om lærerne blitt mer opptatt av å begrunne praksis, med bakgrunn i både forskning og resultater. På den andre siden er de kritiske til et stort resultatfokus og de mer spesifikke verktøyene som iverksettes for å forbedre resultatene. Måtene dette blir gjort på og mulige sosiale effekter av standardisert testing er viktig kunnskap for politikerne.

Teoretiske perspektiver på profesjoner og profesjonalitet kan være hensiktsmessige for å studere ekstern kontroll av lærernes arbeid, og for å belyse mulige tolkninger av hvordan lærerne forholder seg til ansvarliggjøring. Analysene viser hvordan lærernes forhandlinger kommer til uttrykk gjennom det som kan beskrives som diskursivt legitimerings- og grensearbeid. Dette kan tolkes i lys av det jeg beskriver som en 'double-loop' – karakter ved lærerens ansvar for elevenes læring, det vil si at lærerne er ansvarlige for det elevene selv er ansvarlige for. Hvis lærerne opplever at politiske initiativ, som jeg her har undersøkt primært ved å se på nasjonale prøver, fører til at det blir vanskeligere å motivere heller enn å styrke elevenes motivasjon og engasjement, samt at initiativene medfører et snevrere syn på læring, skaper dette dilemmaer for lærere. Hvordan lærere forholder seg til slike spenninger kan tolkes fra et performativt perspektiv, ved at lærere legitimerer hva de gjør eller ikke gjør i klasserommet på bakgrunn av faglig kunnskap og verdier. Det kan også tolkes fra et organisatorisk perspektiv, ved at lærere rekonstruerer den faglige diskursen slik at de er i bedre posisjon til å beholde 'kontrollen' over klasserommet. Ikke minst blir måter å skape relevans og legitimitet på viktig når lærerne må forholde seg til obligatoriske praksiser som nasjonale prøver, men som kan utfordre faglige og personlige verdier. Avhandlingen gir også et metodisk bidrag med hensyn til hvordan analyser av språk kan gi fruktbare og analytiske innganger til på hvilke måter lærere gjennom språket forsøker å ta kontroll over kunnskapsområdet sitt. Jeg har diskutert hvordan

diskursanalyse kan brukes til å undersøke forholdet mellom politikk og praksis, men på en måte som kombinerer deltakernes beskrivelser med teoretiske fortolkninger.

Eli Munkebye: Dialogue for learning – the explorative natural science dialogue in outdoor education

Sammendrag
Avhandlingen er skrevet innenfor fagområdet naturfagdidaktikk med fokus på naturfaglige utforskende dialoger mellom lærer og elever. Studien søker svar på i hvilken grad det forekommer naturfaglige utforskende dialoger når det praktiseres uteskole, hvordan dialogene utvikles samt lærernes valg av strategier for å støtte elevenes læring.

Det er en kvalitativ studie og det som analyseres er naturfaglige utforskende dialogsekvenser mellom lærer og elever. Dialogsekvensene er hentet fra fem lærere som underviser på tredje- og fjerdeklassetrinnet og deres elever, når undervisningen er flyttet fra klasserommet til et nærliggende naturområde. Initiativ- og responsanalyse, samt strategi- og kohesjonsanalyse er sentrale metodologiske tilnærminger.

Studien viste at omfanget av utforskende dialogsekvenser varierte mellom kasusene. Mange dialogsekvenser kan forklares ut fra blant annet lærerrollen i form av en tydelig og fokusert lærer med forventninger til elevenes deltakelse. Dette viser at uteskole kan utgjøre et potensial for utforskende naturfaglige samtaler.

Lærerne styrte utviklingen av dialogen i ulik grad, dialogene var fra sterkt lærer- til elevdominerte. Det ser imidlertid ikke ut til at det viktigste er hvem som dominerer, men at læreren er i stand til å gjøre kloke valg med hensyn til i hvilke situasjoner det er hensiktsmessig å åpne opp dialogen for elevenes stemmer og når det er hensiktsmessig å lukke den for å slippe fram de naturvitenskaplige ideene. Lærerne brukte et bredt spekter av strategier for å få elevene til å bidra i dialogsekvensene, det viste seg imidlertid at elevenes responser var av stor betydning for utvikling av dialogen.
Avhandlingen bidrar til å gi et bilde av hvordan uteskole kan være en arena for utforskende dialoger. Dette avhenger imidlertid av lærernes naturfaglige kompetanse og deres didaktiske refleksjoner.

Studien er utført ved Universitetet i Agder og Høgskolen i Sør-Trøndelag, med tilknytning til Det utdanningsvitenskapelige fakultet, Universitetet i Oslo.

Margrethe Naalslund: Why is algebra so difficult? A study of Norwegian lower secondary students' algebraic proficiency

Abstract
Based on results from the TIMSS studies documenting weak performances in algebra among Norwegian students, this thesis investigates the question "Why is algebra so difficult?" by studying cognitive processes involved in solving algebra problems. *Conceptual understanding, procedural fluency, adaptive reasoning,* and *strategic competence* comprise what in this thesis is called *proficiency,* and serve as a theoretical framework for the empirical data presented and analysed. Written test responses from 412 8th-graders and 417 10th-graders constitute the empirical data, in addition to follow-up interviews with selected students. Diagnostic information, where strategy choices, errors, explanations, and justifications are central elements, is emphasised in the analysis.

The findings suggest a cognitive gap between informal and formal reasoning in grade 8 and pinpoint differences and similarities in formal proficiency in both grades. Formal procedures seem to be used in a highly algorithmic manner, not guided by a deeper understanding of linked core concepts (e.g. equivalence). Many of the same problems were seen in both grades, such as misconceptions due to limited arithmetic knowledge and over-generalizations in the transition from arithmetic to algebra. In addition, an inability to explain and justify their reasoning was evident among the majority of students. An overall message is that algebra teaching could benefit from an increased focus on different aspects constituting algebraic proficiency beyond the "skill versus concept" outlook, and that these aspects are closely intertwined in the development of proficiency.

Guri Anne Nortvedt: Norwegian Grade 8 students' competence in understanding and solving multistep arithmetic word problems

Abstract
This thesis reports on a mixed methods study consisting of two approaches designed to investigate 13-year-old Norwegian students understanding and solving of multistep arithmetic work problems. The process understanding and solving word problems is complex: when working on problems, students need to comprehend the problem text, find an appropriate plan, and execute the identified calculations (Mayer, 2003: Verschaffel, De Corte, & Greer, 2000). Consequently, students need to draw on general reading strategies as well as their mathematical proficiency when solving such problems.

The overall research question of the study reads: "What is the competence of Norwegian grade 8 students in understanding and solving multistep arithmetic words problems?" Competence is a keyword in the research question and refers to

mathematical knowledge and strategies. As competence is a developing aspect of human activity, both what students can do on their own and what they do with support from a more competent other is included in the research topic. In addition, both process and product aspects of problem-solving are of interest. The main aim of the study is to investigate the integrated process of solving multistep arithmetic word problems, with a focus on the following three sub-aspects: (1) the interaction between reading comprehension and mathematical proficiency; (2) students' strategy use when solving multistep arithmetic word problems, as well as causes for difficulties encountered in this process (3) the scaffolding pattern observed when students' word problem solving is scaffolded by a more competent other, as well as relationships between reading comprehension, mathematical proficiency, and scaffolding.

A dual approach was designed in order to achieve the aforementioned in. In Study 1, students (N=19) gave task-based interviews while working on a collection of eight multistate arithmetic work problems. This interview was specifically designed to allow the students to work independently from long intervals and to ensure that the problem-solving process at some point came to a full stop, at which point scaffolding was introduced. The analysis aimed at investigating students strategy use, causes for difficulties and scaffolding patterns.

In Study 2, results of a national sample (N=1,264) on national tests in reading and numeracy were linked at the student level in order to perform the correlation analysis of the relationships between reading comprehension and numeracy in general and word problems in particular. In addition, qualitative analyses of school patterns were used to investigate differences between groups of students**.**

This research study offers methodological theoretical and empirical contributions to the research field of word problems. First, the design of the study is more complex than what was often applied, transgressing the borders between qualitative and quantitative approaches and traditions. The goal of the analysis was to look at not only how well students solve word problems, but also how they cope with these problems. The design provided means for linking process and product data, in addition to enabling analysis of both individual and co-constructed processes, which strengthens the inferences made about students' competence in solving multistep arithmetic word problems. Second, the research was performed with an age group of students and the genre of problems (multistate arithmetic word problems) that are rarely researched.

Overall findings confirm and extend prior findings. In Study 2, reading comprehension was found to have a positive correlation to solving multistep arithmetic work problems, $r =. 631$, $p < .001$, indicating that good readers most likely are also good word problem solvers.

However, it was found in both studies that students at all ability levels make processing errors when reading would problems, such as struggling with relational statements. Many students directly translated keywords. An analysis of answer patterns in both studies indicated that students with above-average numeracy skills and below-average reading skills compensated for more reading comprehension by recognising stereotype word problems to a larger extent than other above-average numeracy students. However, the students more often overused keywords.

In Study 1, below-average numeracy students were found to work with erroneous or simplified situation models for all non-scaffolded word problem solutions. Students needed scaffolding to execute basic operations, including informal methods. These students, to a larger extent, played the words problem game for instance they produced only one solution when the problem clearly stated that more than one solution was to be found. The proficient student in Study 1 primarily required scaffolding to help monitor the process of solving unfamiliar word problems. These students knew when a word problem was not comprehended.

Rolf Vegar Olsen: Achievement tests from an item perspective: an exploration of single item data from the PISA and TIMSS studies, and how such data can inform us about students' knowledge and thinking in science

Abstract

Summary of chapter 1

The thesis was introduced in this chapter by presenting the fundamental rationale for why analysis of items, either one-by-one, or by the study of profiles across a few items, is worthwhile. This rationale was based on a model of how items typically are correlated with each other and to the overall score in an achievement test such as those in TIMSS and PISA. It followed from this model that if we represent the total achievement measure by one overall latent factor, only a small fraction of the variance in the scored items is accounted for by a typical cognitive test score. Furthermore, this argument was brought one step forward by also considering the categorical information in the codes initially used by the markers.

Before the variables in the data file are scored, they are nominal variables with codes reflecting qualitative aspects of students' responses. Taken together with the theoretical model of the scored items, it was concluded that further analysis of the single items would be reasonable, and would involve the analysis of information beyond that contained in the overall score. All the empirical papers in the thesis are based on this rationale: to analyse the surplus information in the items. The purpose of the thesis was then formulated as an exploration into the nature of this surplus information, and the potential of using this information to describe qualitative differences at the student or the country level. Furthermore, the underlying motivation for doing this was stated as a desire to inform the science education community about the potential for, and

limitations of, using the data from LINCAS in secondary research. This latter issue was elaborated and discussed in the next chapter.

Summary of chapter 2

This chapter gave a broad presentation of LINCAS, their policy relevance, and their link, or lack thereof, to the field of science education research. The chapter consisted of several related elements that, taken together, addressed the issue of why and how researchers in science education could or should engage in analyses of LINCAS.

This was done by presenting the historical development of LINCAS, from the first IEA studies by the end of the 1950's to the contemporary studies PISA and TIMSS. I suggested that the development in this period reflects broader societal issues. Moreover, I suggested that the development illustrates a tension or dilemma that LINCAS have been confronted with from the very beginning: LINCAS was initially framed by the idea that international comparisons could be the basis of a powerful design for studying educational issues. Thus, the main idea driving the genesis of LINCAS (which I labelled Purpose I) was an ambition to utilise the international variation in the study of general educational issues. This research base has been maintained throughout the history of LINCAS. What made it possible to conduct the increasingly more expensive studies was the fact that policy makers evaluated the studies as providers of policy-relevant information. Over the years there has been a shift towards the purpose of finding evidence for effective policy at the system or national level (which I labelled Purpose II), and the discussion in this chapter demonstrates that this vision for LINCAS is very visible in the PISA study. It would be fair to say that my thesis aims to promote Purpose I, and, furthermore, it aims to promote the view that the tension that is often perceived between the two purposes is to some degree based on a lack of communication and interaction between the policy makers and the educational researchers. The chapter then turned to a comparison between PISA and TIMSS. This is an issue that in itself is worthwhile because there are some indications that users of the information may be confused by discrepant results in the two surveys. However, by examining the differences between the studies, it is evident that the results should not be compared in a simplistic manner: they have different designs targeting different populations and different levels of the school systems, they have defined the achievement measures differently, and even if many countries participate in both studies the composition of the countries in the two studies is clearly not the same. Chapter 2 continued by discussing how science education may be linked to the policy context by engaging in secondary analysis of data and documents from LINCAS. This was not to argue that all, or even most, of the research in science education should be linked to PISA or TIMSS. Nevertheless, a relatively comprehensive review of possibilities for secondary analysis related to LINCAS was presented in the chapter, and the increased potential for such analyses relating to scientific literacy in PISA after the 2006 study was emphasised.

Summary of chapter 3

Chapter 3 gave an overview of some methodological issues that have heavily influenced my work. It began by placing my work in a tradition that could best be labelled as exploratory data analysis. The main idea of this tradition is that when confronted with a data set we should seek to develop a description of the overall structure in the data, the multivariate relationship, which is a challenging task since there is no general procedure to follow for finding such overall patterns in the data. In addition the general issue of the nature of the information in the cognitive items in TIMSS/PISA was explored in this chapter. A novel innovation in TIMSS was the double digit codes and the associated marking rubrics used for the constructed response items. With TIMSS it was acknowledged that using only multiple choice items, which before TIMSS was commonplace in most large-scale assessments, would seriously limit the range of competencies activated by a test. By using open-ended questions, giving students the opportunity to construct their own responses, TIMSS had the ambition of developing descriptions of how students' represented and made use of concepts in science. The double digit codes were used to preserve that information. This was also the idea in the science assessment of PISA 2000, although the generic system was slightly modified. However, with PISA 2003, and with the items that have undergone field trials before PISA 2006, it is evident that the use of such coding is gradually disappearing. The reason for this change is not entirely clear, but it may be suggested that the codes have been of little use internationally. Nevertheless, constructed response items will still be used since they allow for the testing of competencies other than the selected response formats. The paradoxical consequence of this is that from PISA 2006 more information about students' thinking and knowledge will be available from analysis of the multiple choice items than from students' own written accounts of their reasoning and thinking, since the former at least include a code reflecting the response selected by the students. The constructed response items that were originally introduced into these assessments as tools for making the students demonstrate their thinking and reasoning are, in the marking guides for PISA 2006, more or less directly reduced to a description of how to score the items. Even if the marking guide includes explicit descriptions of the criteria for scoring, for the great majority of items there are no longer separate codes for students with different types of responses. I will suggest that this development was perhaps inevitable given that these codes were not extensively used or reported on in the international reports. However, I regard this development as a decrease in the potential for communicating how students typically think and interact with the items in tests like PISA. Furthermore, this development can be viewed as unfortunate from the perspective that such data could possibly be an important resource for secondary analysis aimed at studying students' understanding of very specific scientific concepts or phenomena. Figure 3.3 provided some bipolar characteristics of analyses of information at different levels in item processing from specific written responses, through the coded responses, and finally to the scored items. Information is continuously and consciously peeled off in this process. In the first process of coding, all aspects that are seen as irrelevant for the overall intention of the response are peeled

off. This may, for instance, be information regarding errors in spelling, errors in grammar, and other very specific elements in the response. However, it may also be information that reflects characteristic features of students' thinking and knowledge. The marking guide has to be understood similarly by all markers, in all countries, and thus, it is a necessary condition that the number of codes are limited, and that they reflect clearly identifiable features of students' responses. The codes therefore represent classes of typical responses that may be distinguished from each other. In the next process, when the items are scored, all aspects other than the overall quality or correctness of the item are peeled off. The score can therefore be considered as not representing aspects of the responses as such, but rather as representing aspects of the ability that students have used to create their responses. At least this is the idea. However, as demonstrated in Figure 1.1, the score information at the single item level is still highly specific for the item. Furthermore, chapter 3 addressed more specifically the methods used in one of the papers: correspondence and homogeneity analysis. I have so far not seen any other analysis where these, or similar tools, are used to study the relationship between nominally measured cognitive variables. In that sense the work undertaken in this paper represents an innovative approach to the analysis of data from cognitive tests. The aim of this section in chapter 3 was to write about the methods at a level requiring very little mathematics. This was a conscious choice in order to make this part of the text available to a more diverse group of readers. One consequence of this would be that interesting aspects of the methods are not commented on. Furthermore, since the language of mathematics is a useful tool that allows for very precise and unequivocal communication, another unfortunate consequence may be that the text is ambiguous, thus allowing misunderstandings to develop. Nevertheless, writing for a wider audience has forced me to challenge my own understanding of the methods I have applied.

Tine Prøitz: Conceptualisations of learning outcomes in education –and explorative cross-case analysis of policymakers, teachers and scholars

This article-based thesis presents the findings of an exploratory qualitative multiple case study of how learning outcomes are conceptualised in education policy. Learning outcomes is considered a key concept in a changing education landscape. International organisations with influence on national education policy present one commonly shared understanding of learning outcomes. However, the review of outcomes literature in this study illustrates that the concept is contested in research, with a debate deeply rooted in issues of what constitutes learning and how it can be valued. This contradiction between different understandings of the concept of learning outcomes calls for a closer look at how the concept is understood in education. While contradictory concepts in education are not unusual, in this case it seems important to explore what understandings are at play to avoid taking them for granted and enable an informed and open debate of what should be valued and appreciated as learning.

The multiple case study consists of three individual case studies each presented in one paper. Study 1 illustrates how learning outcomes is contested among scholars. Two broad clusters of conceptualisations of the concept have been identified. In the established cluster learning outcomes is considered as results-oriented, full ended and measurable. In the alternative cluster, learning outcomes is understood as process-oriented, open-ended and with limited measurability.

The majority of the scholars studied consider learning outcomes as a concept for the purpose of educational, instructional planning and curriculum development – thus with an internal focus. The study illustrates how several conceptualisations are at play within academia.

Study 2 illustrates how teachers conceptualise learning outcomes when describing their grading practices. At an overall level teachers report to understand learning outcomes in compliance with the national curriculum and regulations for grading of the outcome based reform of 2006. It also displays a tension between characteristics of school subjects and universal regulations for grading. The study suggests that in the eyes of teachers different subjects have different degrees of challenges in fulfilling government recommendations and universal regulations for grading within an outcomes based system, some being more easily adaptable than others.

Study 3 explores how Norwegian policymakers conceptualise learning outcomes during a period of 14 years. The study illustrates how policymakers in subsequent governments have embraced the concept. It also illustrates that the concept is not a controversial policy issue in itself and that policymakers understand it as results-oriented, full-ended and measurable. Policymakers use the concept in relation to the external purpose of accountability. However, the study also suggests that under a consistent learning outcomes umbrella, governments introduce a variety of often contradictory policy initiatives, eg monitoring of outcomes for decentralisation and local accountability vs. monitoring of outcomes for more central state control, possibly downplaying local accountability.

The cross-case analysis of the three studies suggests that there is an overall established and dominant understanding of learning outcomes. It also identifies alternative approaches to the concept presented by scholars and teachers. The analysis illustrates how several understandings are at play in practical language use between and within groups of actors. The study proposes a four-part model for the identification of conceptualisations of learning outcomes. The findings of the study indicate that a dominant conceptualisation limits the understanding of learning outcomes while other available understandings are seemingly left unexplored by the actors studied.

The study contributes methodologically to the field of learning outcomes by studying learning outcomes as conceptualised by the speech acts of three groups of actors. It

contributes theoretically to the field by presenting a theory-based analytical framework, which over the course of the study advances into an empirically grounded four-part model for conceptualization of learning outcomes. The study is relevant in the way it offers a model for consideration of different approaches to learning outcomes in education, and in its potential for identification of practices that manage to balance external requirements of policymakers with internal requirements of education.

Gunnhild Tveit Randen: Adequate skills in Norwegian? Mapping of immigrant pupils' language skills

Abstract

The goal of this study is to investigate the language proficiency of five L1 Russian-speaking first-grade students learning Norwegian as a second language in school, and the use of different forms of assessment. I also discuss how the language proficiency of minority students can be assessed in school. This gives the study two objects of research: the students' language proficiency and the schools' assessment practices. The Norwegian law of education ensures language minority children's rights to adapted education in Norwegian until they are sufficiently proficient in Norwegian to attend mainstream schooling. This gives every child individual rights according to his or her level of proficiency in Norwegian, which makes the operationalisation of language proficiency a significant issue in assessment in school. The study answers two research questions concerning language proficiency and two questions concerning assessment. In terms of language proficiency, the research questions concern how minority students' language proficiency can be assessed in light of the school's chosen education program, and which correlations there are between L1 proficiency and proficiency in Norwegian. As regards assessment, the main question relates to how an assessment test made for L1-Norwegian students will work for language-minority students, and, as a followup question, how students' language proficiency can be assessed through conversation.

Two theoretical strains are of particular interest: theories of communicative language ability and theories of language awareness. When language ability is assessed in the first grade, language awareness is a central part of the assessment, because language awareness is considered a key aspect for the development of literacy skills. Yet language awareness is often not mentioned in models of language use. In the theory section an alternative model for description of language proficiency is presented, which includes language awareness and therefore provides a good basis for assessment of first-grade students' language proficiency.

The study has a qualitative approach, i.e. case studies. The data consist of the case students' oral storytelling, a standardized test, conversations about pictures, examples of students' reading and writing, interviews with students and teachers, external teachers' assessments and observations.

The analysis has two main focuses: description and assessment of language awareness, and description and assessment of text proficiency. Language awareness is investigated by means of a standardized test. According to the test score the students have a low degree of language awareness, which would normally indicate that the students have difficulties in acquiring the skills of reading and writing. This result is not, however, consistent with the assessment of literacy skills, which shows that four of the five students can already read and write at an age appropriate level in two languages and using two alphabets. A qualitative analysis of the conversation in the test situation indicates that the low test score is caused by other elements than a low degree of language awareness, and that this test is not useful for assessing these students' language awareness.

Text proficiency is studied by investigating the students' retellings of stories in Norwegian and Russian. Analyses of narrative proficiency and cohesion in the stories show great variation in each student's proficiency in the two languages. The students' opportunity to demonstrate their text proficiency in a certain language is dependent on some degree of mastery of the linguistic structures of that language. Assessment of text proficiency in the language the student knows best gives an indication of the student's actual proficiency, independent of language proficiency. This has implications for teachers' adaptions in class. In the study of text proficiency, attention is also directed towards the adults' scaffolding during the students' retellings. The analyses illustrate that the adults adjust their language in interaction according to each student's needs. Categorisation of the adults' utterances indicates which language elements the students need to develop, and therefore clarifies the needs for individual adjustments.

The final chapter of the dissertation comprises a discussion of which forms of assessment may be useful for the assessment of language-minority students' proficiency in Norwegian.

Knut Roald: Quality Assessment as organisational learning between school and school owner

Abstract

This thesis focuses on how schools and school owners collaborate on questions of quality assurance in a national system based on management by objectives, performance management and accountability. Norwegian schools face major challenges both in terms of academic achievement, social equity and increased dropout rates during the 13-year basic education. To meet these challenges, Norway follows international management trends placing increased emphasis on quality assessment and local responsibility. Municipalities are – as school owners – held accountable while simultaneously expected to attain national as well as local targets by working as learning organisations. Academic theories on organisational learning thus become a relevant framework for studying how school and municipal levels collaborate on the

quality assessment work. This study is part of a larger research project financed by the Norwegian Research Council, "Achieving School Accountability in Practice" (ASAP), involving analyses of the quality assessment work carried out at the national, municipal and institutional levels in Norwegian schools.

My work has been linked to a subproject within ASAP aiming at achieving research based knowledge about how local actors understand the extended room to maneuver they have been granted through the latest reforms. The objective of my study has been to develop knowledge on the organisational learning processes that take place between schools and school owners when they collaborate on issues of quality assessment. The empirical material builds on focus group interviews and individual interviews in municipalities that have over time been engaged with developing quality assessment systems. To establish a deeper understanding of learning processes between the school and municipal levels, I have studied the challenges and possibilities as they are perceived by academic, administrative and political leaders in their work on assessing quality in schools.

The theoretical framework for the thesis consists of different approaches to organisational learning, and draws on the work of Gregory Bateson, James March, Argyris & Schön, Peter Senge and Nonaka & Takeuchi. Different perspectives on learning in organisations are discussed along with central aspects of the criticism raised against the discourse on organisational learning. The thesis also looks at the relationship between organisation, administration and learning to clarify different academic views on quality and assessment. The study shows that the school's quality discourse and the day-to-day teaching and learning work can go on in two independent spheres. Results from quality assessments are passed on to politicians, managers, teachers, students and parents, but the daily work in the classroom tend to continue much as before. At the same time the study shows that municipal managers are aware of this challenge and work to develop strategies to stimulate a knowledge-developing quality work between the users, the groups of professionals, the administration and the politicians. The transition from information to knowledge is challenging schools and school owners alike. National, municipal and school-based programs for quality assessment generate large quantities of information, but as the study shows, the assessment information in itself does not lead to new understanding or active development work. Unless data is presented in ways that provide collective insights and commitment, increased availability of information can actually be counterproductive.

One of the findings in this work is that quality discussions both at the school and municipal levels are often characterised by incommensurable contributions, ranging from purely positivistic ones concerned with causal explanations, to social-constructivist approaches that view quality work as intersubjective meaning making. The various actors engage in different discourses. In order to establish a knowledge-developing quality assessment procedure between schools and school owners, some

basic prerequisites appear to be: - a broad body of qualitative and quantitative information - use of both internal and external quality assessment - development of joint arenas and meeting forms conducive to productive work on quality assessment

The study shows that assessment work at one level is dependent on good assessment processes at other levels of the educational system. High quality formative and summative pupil assessment creates good conditions for school-based assessment, which in turn increases the school's ability to make use of external assessment. Similarly, constructive collaboration between teams of teachers, trade unions and headmasters at individual schools seems to have a positive effect on the interaction between the group of headmasters, the municipal administration and the politicians. The quality assessment work thus seems to function productively when an assessment culture is created from below throughout the municipal school system. Such an assessment culture also seems to be characterised by balance and coherence between the experience–expectation, reflection–action, structure–culture and individual–collective axes. In order to build this kind of bottom-up culture for assessment, the top leaders have to understand this relational way of thinking. No assessment tools and methods stand out as better than others. The crucial factors seem to be the provision of ample assessment information and high-quality discussion and interaction processes in the quality work. When schools and municipalities hire external evaluators, the study shows that it is decisive to cooperate on identifying the areas, criteria and methods of assessment. Quality assessment work is unlikely to be productive if teachers, students and parents are merely informers and suppliers of data. They must also be involved in determining what is to be examined and what changes are necessary.

The study shows marked divisions between unsystematic, systematic and systemic features of quality work. A systematic approach emphasises quality systems, with linear plans for the use of tests, user surveys and information on outcomes. The systemic approach places more emphasis on quality work, including descriptions of dialogue and interaction across traditional hierarchical levels to identify knowledge developing processes. Both groups of professionals, the administration and the politicians must be involved as co-creating participants in the quality work. This enables them to combine control-oriented, decision-oriented, teaching-oriented and process oriented strategies. The management of knowledge-developing processes seems to require forms of preparation, implementation and follow-up that differ from traditional bureaucratic working methods. The main task of managers is not primarily one of preparing the relevant information and suggesting decisions, but one of preparing problems for professional discussion and suggesting distinct process steps that can mobilise a high degree of reflection, creativity and responsibility between those involved in the quality work. Actively using systemic patterns of interaction, schools and school owners can establish a genuine, local room of maneuver for the development of the school's teaching and learning efforts. Deprived of this insight, municipalities could perceive the logic of accountability as implying that quality assessment work increases the distance

between schools and control-oriented municipal bureaucrats and politicians. Based on this observation, the thesis also raises the question of how the state level could help municipalities become school owners with an advanced capability for systemic development.

Lise Vikan Sandvik: From goals to meaning: a study of writing and assessment culture in German undergraduate teaching

Abstract

The overriding aim of the study is to contribute to an increased understanding of the link between assessment and writing in foreign language teaching. There is a particular focus on assessment as a learning enhancing tool in relation to the students' development of their writing skills. This study has been conducted from both student and teacher perspectives, but the teacher's assessment skills and the significance of this in the student's learning process is particularly prominent.

The teaching of writing in foreign languages in Norway principally occurs in the school context. When the teacher has to give feedback on students' texts, she is faced with many choices. How she chooses to respond to the texts has implications for how the students experience their own work in order to create meaning in a foreign language and for how the relationship between the students and the teacher is perceived. This context is the subject of this study. I obtained the data material, which consists of observations, interviews, students' texts and questionnaires, in a junior secondary school observing one teacher and her students. The data was collected over a period of 18 months.

One of the principal findings of the study is that formative assessment as a mediating artefact is significant to the students' writing process. Provided that the students have an understanding of the purpose of the writing and the aim of the feedback they receive during the writing process, they will also work on the text more thoroughly on all levels in the next phase of writing. A shared understanding of the aims of the writing also contributes to the learning process becoming more transparent, opening up for a better collaboration on the writing process among the students.

Another important finding of this study is that it appears that the teacher's assessment ability plays a central role in the students' development of writing skills in German. The whole assessment culture in the classroom appears to develop when teacher and students have a common reference of assessment and when the process, purpose and context is clear and understandable to all the participants while building knowledge.

Guri Skedsmo: School Governing in Transition. Perspectives, Purposes and Perceptions of Evaluation Policy

Abstract

This thesis focuses on perspectives, purposes and perceptions of evaluation policy. Based on conceptual and documentary analysis, I explore how national school authorities develop policy instruments to regulate and renew comprehensive education. A national survey directed to school principals in Norway in 2005 provides data for the empirical study. By comparing purposes of evaluation policy and responses from the school principals, I examine how school governing partly transforms as a consequence of the new orientation. One of the main elements is the introduction of a national comprehensive evaluation system during the early 2000's. This system consists of different evaluation tools which create new governing structures, processes, as well as expectations to schools and principals. The question is in what way and to what extent school governing is transformed by this new evaluation policy.

Empirically, the study is restricted to recent changes in educational policy and school governing in Norway. I investigate how governing structures and processes develop and transform through the introduction of a national comprehensive evaluation system in Norwegian education. The focus of the thesis is directed towards theoretical perspectives, policy purposes and key actors' perceptions of evaluation policy. I aim to develop analytical concepts to better understand the rationales and the complexity in educational policy which have consequences for school governing. In addition, I explore the linkage between national evaluation policy and principals' perceptions of evaluation tools and new accountability forms.

My analysis shows that national evaluation system is characterised by output regulation. This implies that information about educational outcomes is used as a foundation for school governing. In spite of the policy discourse related to the national evaluation system, the governing modes of the different evaluation tools imply first of all a focus on measurement of student achievements. Second, comparisons of results are a central element in gathering information about educational outcomes to gain oversight. The results of groups of students, schools, municipalities are compared, as well as the national progress seen in relation to outcomes produced in other countries.

Although the arguments about establishing a national evaluation system do not change much from the early 1990's until 2004, the language used in policy documents shows an increasing output orientation in the policy texts. From focusing on quality development in education, there is a shift in the early 2000's towards emphasising educational outcomes in terms of student achievements, and how to improve outcomes.

The increased focus on outcomes in education policies is to a certain degree also reflected in principals' perceptions. The analysis of the survey data showed that the

national evaluation system has led to some changes in the school practice in terms of giving priority to areas such as routines to following up the results achieved, the use of specific programs or methods to improve the students' basic competencies, and the work on staff development. In addition, according to the principals' perceptions, the frequency of testing in schools has increased.

Based on the analysis, some problematic aspects of the formulated policy purposes related to the tools' modes of regulations, but also to inconsistencies between the purposes and how the tools are perceived by the principals are discussed. Furthermore, I argue that output regulation as a new way of coordinating the education system, will influence on the relationships between national and local authorities as well as the schools, and create new patterns of interactions.

Sidsel Skjelten: The search for writing quality differences in pupil texts. What differentiates very good texts from good texts?

Sammendrag
Elevane sin skrivekompetanse i norsk ved avslutta grunnskule er interessefeltet i avhandlinga. Nærare bestemt er det forska på kvalitetsforskjellar mellom gode og middels gode tekstar ut frå korleis dei er vurderte til eksamen i 10. klasse. Det er såleis grenseoppgangen mellom karakterane 4 og 5 som har stått i fokus, og målet har vore å fange opp kvalitetsforskjellar mellom dei to karakternivåa ut frå ei mest muleg heilskapleg tilnærming til tekst. Forsking på elevtekstar og elevanes skrivekompetanse i skulekonteksten knyter prosjektet til praksisfeltet og til språket i bruk. Språkbruksperspektivet gjer at avhandlinga kan plasserast innanfor vitskapsdisiplinen anvendt språkvitskap, men at det praksisnære blir kombinert med teoriutvikling.

Materialet er henta frå det nasjonale KAL-korpuset. Det gir tilgang på elevtekstar i ulike kvalitetskategoriar, i tillegg til utfylte vurderingsskjema frå sensorane. Studien er kvalitativ og teksttolkande, der det handlar om å tolke meining i to ulike typar tekstar: 1) utfylte vurderingsskjema og 2) elevtekstar. Det unike med materialet er at kvar elevtekst er vurdert av tre sensorar i tillegg til dei to ordinære, noko som styrkjer gyldigheita i forskingsopplegget.

Studien viser at *tekststruktur, truverd, engasjement* og *rettskriving* utgjer kvalitetsforskjellar mellom dei to karakternivåa. Gjennom abduktive tilnærmingar er kriteria operasjonaliserte i møte med elevtekstane, og etablert teori frå språk- og tekstvitskapen er henta inn for å forklare mønster og strukturar i tekstane, hovudsakleg frå retorikk og systemisk-funksjonell lingvistikk. På den måten er det konstruert nye teoretiske modellar som kan seie noko om den typen kvalitetsforskjellar som prosjektet har jakta på. Avslutningsvis er dei fire kriteria plasserte inn i eit nyutvikla vurderingsskjema for å synleggjere kvalitetsforskjellar som er komne fram i analysane.

Prosjektets bidrag kan knytast til to område: 1) til skulekonteksten, der vurderingskriterium utgjer eit viktig pedagogisk verktøy i arbeid med elevanes skriveutvikling, og 2) til forskingsfelt som er opptekne av elevanes læringsutbytte, tekstkompetanse og kriterieproblematikk.

Abstract
The focus of this dissertation is on pupils' writing skills in Norwegian at the end of lower secondary school (the 10th school year). More precisely, this research covers writing quality differences between very good texts and good texts based on 10th grade exam assessments. Thus, the focus is on the differentiation between the marks 4 and 5 on the grading scale, and the goal has been to find the writing quality differences that exist between these two levels, based on an approximation of the entire text. Research on pupil texts and writing competence within a school context, connects this project to language in use. The language in use perspective allows the dissertation to be placed within the science of applied linguistics, whereas the practical aspect is connected to theory development.

Research material is collected from the Norwegian national KAL-corpus, which provides access to pupil texts within various quality categories, in addition to completed assessment forms from the censors. The study is both qualitative and text interpretive, as two different text types are interpreted: 1) completed assessment forms, and 2) pupil texts. The material is unique as the pupil texts have been assessed by three censors, in addition to the two ordinary censors, a fact that strengthens the validity of the research project.

The study indicates that *text structure, credibility, engagement and orthography* make up the quality differences between the two grading levels. Using abductive approaches, the criteria are operationalised in connection with the pupil texts, and established theory within language and text science has been brought in to explain patterns and structures in the texts, mainly based on rhetoric and systemic functional linguistics. Thus, new theoretical models have been construed to explain the type of writing quality differences found in this research. Finally, the four criteria are placed in a recently developed assessment form to visualize writing quality differences resulting from the analyses.

The project can contribute to two areas: 1) in a school context, where the assessment form is a significant pedagogical tool for developing pupils' writing skills, and 2) within the research field of pupils' learning outcomes, text competence and the problems connected to criteria.

This research project has been carried out at the Faculty of Educational Sciences, University of Oslo. The location throughout the research period has been the University of Stavanger.

Ola Johan Sjøbakken: Frequent teacher–pupil dialogue as in an action research perspective

Abstract

By utilising action research as a means of studying teacher practice and reflection this research seeks to provide new knowledge of regular, formalized teacher-pupil dialogues. Two research questions have been posed. First, *what is the importance of pupil feedback in regular, formalized teacher-pupil dialogues?* is based on international research on feedback and is grounded in an alternative approach to the analysis of teachers' assessment of feedback to promote student learning. The second: *How can a research partnership in a longitudinal study provide knowledge of the regular, formalized teacher-pupil dialogue?* is explored by drawing upon a theoretical and methodological approach to action research, the analysis of pupil plan books and letters from teachers.

The findings in this study are that the systematic teacher-pupil dialogue influences the teacher's way of thinking; it leads to a greater awareness of adapted education, and provides insight into how the pupil conceptualises and structures his own knowledge. Teachers in this study described dilemmas such as lack of time, being squeezed between a focus on the subject and addressing the social needs of pupils, and the difficulties of carrying out the teacher-pupil dialogue. The study also reveals openness and uncertainty as part of the process of teacher feedback, and occasionally metaphors are drawn upon because language is insufficient in describing the various phenomena they encounter in their work with regular, formalized pupil conversations. On the other hand, they note the potential in the tool itself; for intimate conversations with the pupil and enhancing co-operation with the home. They don't experience the formalized teacher-pupil dialogue as lacking in complexity. On the contrary, it represents a way of working that also includes uncomfortable choices and places significant demands on the individual. Despite different forms of practice, feedback is the single factor that takes up the most time and place in the dialogue. A new model is developed for the understanding of feedback as an alternative to the more established four-part typology of feedback: task, process, self-regulation and personal level. This is one of the study's most important contributions.

It is *longitudinal* study where a group of teachers were followed over a period of in total three years, and the period in-between can be characterised as a form of *self-motivated action research*. The Government directives and national policy on systematic, regular dialogue in primary and secondary schooling have been in place from 2009, and the increasing focus upon feedback in international research and in national white papers, actualise the need for research on feedback and the formalized teacher-pupil dialogue. Hedmark University College and the Institute of Educational Research, University of Oslo have supported this research and the student has been registered as a doctoral student at the last mentioned institution.

Oddny Judith Solheim: Assessment of reading performance, fundamental conditions for measuring reading

Abstract

During the last 50 years the field of reading research has seen radical changes in the way reading is conceptualized. New theoretical perspectives strongly influence the way reading is measured. There has, however, been a delay in the implementation of these theoretical perspectives in test development. The tendency to confront different theoretical perspectives on reading has recently been replaced by a willingness to integrate different perspectives. This has resulted in a situation where several core concepts *(for instance literacy, depth of meaning, motivation and engagement in reading)* are unsettled or ascribed different meanings by researchers. When it comes to assessment, this situation demands close attention from both researchers and test constructors. If hidden, implicit or unsettled assumptions form the basis for development of assessment instruments a veil is drawn between theory and empirical results. This can result in poor quality of the inferences made on the basis of test scores. In this thesis challenges for measurement of reading, in a time with integration of perspectives are identified and explored. The thesis consists of three studies that trace implications of integrating theoretical perspectives on the development and validation of reading measures.

The focus of study 1 and 2 was on item format, a recurring subject in research and discussion about the assessment of reading. Multiple-choice (MC) is simple and economical in terms of testing time and scoring costs. In relation to large-scale assessments, it has often been discussed whether the use of constructed-response (CR) items may be justified despite the extra expense involved. The decisive factor has been "value added": the extent to which the inclusion of CR items improves our assessment of reading comprehension in a valid and reliable way. The main aim of the first paper was to explore how the operationalisation of depth of understanding in the Progress in International Reading Study (PIRLS) corresponds to the description of reading literacy given in the PIRLS framework. In PIRLS, the CR format is the bearer of central notions included in the theoretical foundation of the test, and this determines how the CR items should be designed and scored to be compatible with the underlying theory of reading comprehension. The study focused on the scoring guides and the relationship between these and the text and items. In the second paper the main aim was to explore the interaction between item format and motivation. Specifically it focused on how motivation contributes differently to reading-comprehension scores depending on item format; and whether students with different levels of reading motivation profit differently depending on item format. The aim of the third study was to explore possible uses of eye-tracking methodology in process-oriented reading test validation. The challenges researchers face in selecting empirical indicators of reading comprehension were highlighted. Research questions where eye tracking methodology may support validation of assessment methods were discussed. Results from a small scale eye-

tracking study where students read and answered questions about a multimodal text were used as example material.

Different data sources were used. The first study was a critical analysis of the relationship between the PIRLS Framework 2001 and the PIRLS Revised Scoring Guides 2001. The study also draws on a representative sample of Norwegian 10 year olds (the Norwegian sample in PIRLS 2001). The sample consisted of 815 pupils. The second paper was an empirical study on data collected in 2008. A total of 217 5th graders from 12 classes in 5 schools participated. The third paper was a theoretical and methodological discussion based on eye-tracking data from a small scale eye-tracking study. The eye-tracking data were collected on 20 pupils from the 7th grade.

The results from the first paper revealed a tendency in the PIRLS assessment towards projecting the depth dimension onto a superficial plane in that a number of observations were accepted as an expression of a qualitative entity. This led to the quantification of a phenomenon originally defined as qualitative. The desire to reduce error of measurement tended to undermine the validation arguments which link results to the theoretical description of the construct. In the second paper, results showed that, after controlling for word-reading ability, listening comprehension and non-verbal abilities, reading self-efficacy was a significant positive predictor of reading-comprehension. For pupils with low self-efficacy in reading, this was a significant positive predictor of multiple-choice comprehension scores but not of constructed-response comprehension scores. For students with high self-efficacy in reading, this did not account for additional variance in either item format. These findings have implications for development of reading assessment instruments and the conclusions we draw on the basis of test scores. Results from the third study indicated that reading behaviour associated with a high level of reading comprehension, may be understood by comparing the pupils' first reading and their reading while answering questions. This has important methodological consequences for research on assessment of reading comprehension.

Are Turmo: Science education and large-scale international comparative achievement studies: A thesis for the degree of Doctor Scientiarum

Abstract

How can large-scale international comparative achievement studies (LINCAS) be used within science education as a scientific discipline? The thesis addresses this question by theoretical discussions and by analysis of data from the PISA study (Programme for International Student Assessment). This study was initiated by the OECD and focuses on 15 year olds' scientific, mathematical and reading literacy. The definitions of the domains focus on competencies that are seen as important for participation as a reflective and concerned citizen in a democratic society. The empirical analysis in the thesis focuses on how data from the PISA study can increase our understanding of

aspects of “scientific literacy”. A key message in the thesis is that LINCAS have a large unused potential for science education researchers. Despite this large potential, it is also argued that the design of such studies can be modified to enhance the possibilities. In particular, it should be possible to construct tests that measure one or a few cognitive traits with high validity and reliability, and also can be used in diagnostic analysis of students’ understanding of fundamental science concepts.

David Wheat: The Feedback Method. A System Dynamics Approach to teaching microeconomics

Abstract

This thesis documents a method for improving undergraduate instruction in macroeconomics. Called the feedback method, it enables students to learn about dynamic behavior in a market economy by using feedback loop diagrams and interactive computer simulation models instead of static graphs or differential equations. There are at least two types of pedagogical problems associated with graphical representation of the economy. First, students seem to have difficulty interpreting static graphs used to illustrate dynamics, which raises questions about the value added by graphs to student understanding. Secondly, the most prominent graph in modern macroeconomics principles textbooks—the aggregate supply and demand (AS/AD) model—appears to misrepresent disequilibrium conditions in the economy and cause students who understand the graph to misunderstand important behavior in the economy. The feedback method emphasizes dynamics rather than static equilibrium conditions. How the economy changes over time in different contexts is the behavioral question that students repeatedly encounter. The structure of the economy is explained in terms of reinforcing and counteracting feedback loops. Student understanding of the source of dynamic economic behavior requires seeking, identifying, and explaining relevant feedback structure in an economic system. Interactive computer simulation activities reinforce the insights gained from studying feedback loops. Even small-scale student participation in model-building seems to facilitate understanding of a larger model; moreover, such participation may build respect for the scientific method and an appreciation for theory building by economists. The feedback method is a structural explanation of economic behavior, but it also provides an improved learning structure for students, and the thesis reports on four experiments designed to test that claim. Two experiments examined student preferences for methods of learning macroeconomics; for example, using static graphs or a feedback loop diagram. The experimental designs were quite different, but the results were the same—a significant majority preferred the feedback method. The most commonly cited reason: feedback loops enable the students to visualize a process in the economy. The third and fourth experiments addressed the performance question. In the third experiment, students showed more understanding of GDP when they had access to a stock-and-flow feedback diagram of the economy. In the final experiment, students using feedback loop diagrams displayed more understanding of business cycle dynamics than other students who had access to an AS/AD graph.

Teaching students to search for feedback structure in the economy and using computer simulation to connect structure with behavior appears to be a promising method for teaching macroeconomics.

Anne Line Wittek: Portfolio as an artefact for learning in higher education. Structures, cultural practice and trajectories of participation

Abstract

The main research question is: what does it mean to implement portfolio as a pedagogical tool in higher education, and how can the portfolio as an artefact structure learning activities and personal trajectories of learning? The study takes its departure in a socio-cultural approach to learning and knowledge. Dialogism is the main focus within the approach. Mediated action is the unit of analysis. The portfolio is conceptualised and analysed as an artefact that is involved in interaction both as a tool and as representations of the tool.

In light of the chosen theoretical perspective, I suggest three forms of representations of the portfolio as an artefact, first as a pedagogical tool, second as a set of structures for activity and third as a set of structures for new knowledge. The research literature within the field is characterised by an idealistic outlook on portfolio as a pedagogical tool, and it is well documented that it has a great potential for student learning. However, the literature says very little about how the portfolio is being constituted in practice and what actually happens when students starts working on their portfolio
The present study investigates how the portfolio as an artefact is constituted within a specific context. The focus is on how students negotiate and develop structures for collective and individual trajectories of participation. The context for the study is the part-time nursing program, and the study is designed as a rich ethnographic study, with a special focus on interaction and application of portfolio as an artefact. Practice in two selected study groups is being analysed both as chronological stories through a year of study, and as more detailed analysis of activity in selected episodes of activity. Individual learning is being understood as personal trajectories of participation, and those will be analysed in light of collective trajectories of participation.

The most important conclusion to be drawn from this study is that portfolio as a structure for activity is constituted in activity as a result of contextual negotiation and interaction. The act of writing in itself is the most important dimension of the artefact and is particularly significant in the construction of learning and activity. However, the analysis shows that the two groups develop very different patterns of interaction. The study draws a nuanced picture of interaction and trajectories of participation. It emphasised particularly the social and contextual assumption for learning and the relation between the collective and the individual dimensions of learning.

References

Aasen, P et al. (2012) *Kunnskapsløftet som styringsreform – et løft eller et løfte? Forvaltningsnivåenes og institusjonenes rolle i implementeringen av reformen*. Rapport 20/2012. Oslo, NIFU/Universitetet i Oslo

Abu-Alhija, F N (2007) Large-scale testing: benefits and pitfalls, *Studies in Educational Evaluation*, 33 (1), 50–68

Adams R J (2003) Response to 'Cautions on OECD's recent education survey (PISA)', *Oxford Review of Education*, 29 (3), 377–389

AERA, APA & NCME (1999) *Standards for Educational and Psychological Testing*. Washington, DC

Afonso N & Costa E (2009) The influence of the Programme for International Student Assessment (PISA) on policy decision in Portugal: the education policies of the 17th Portuguese Constitutional Government, *Educational Sciences Journal*, 10, 53–64, http://repositorio.ul.pt/bitstream/10451/5667/2/Afonso%20%26%20Costa%20(2009).%20ENG%20D4.pdf

Ahmed, A & Pollitt, A (2007) Improving the quality of contextualized questions: an experimental investigation of focus, *Assessment in Education: Principles, Policy and Practice*, 14 (2), 201–232

Allal, L (2005) Assessment and the Regulation of Learning, *International Encyclopedia of Education*, 3, 348–352

APA, AERA & NCME (1954) Technical recommendations for psychological tests and diagnostic techniques, *Psychological Bulletin*, 51 (2), 1–38

Anastasi, A (1976) *Psychological Testing* (Fourth edition). New York, MacMillan Publishing Co. Inc

Anderson, L W & Krathwohl, D R (eds.) (2001) *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives* (Complete edition). New York, Longman

Anderson, J R, Matessa, M & Lebiere, C (1997) ACT-R: A theory of higher level cognition and its relation to visual attention, *Human-Computer Interaction*, 12 (4), 439–462

Andrich, D (2003) On the distribution of measurements in units that are not arbitrary, *Social Science Information*, 42 (4), 557–589

Andrich, D (2004) Controversy and the Rasch Model: A Characteristic of Incompatible Paradigms? *Medical Care,* 42 (1), 17–116, January Supplement: Applications of Rasch Analysis in Health Care. Lippincott Williams and Wilkins.

Arffman I (2010) Equivalence of translations in international reading literacy studies, *Scandinavian Journal of Educational Research*, 54 (1), 37–59

Assessment Reform Group (2002) *Assessment for Learning: 10* Principles. Research-Based Principles to Guide Classroom Practice. Assessment Reform Group.

Assessment Reform Group (1999) *Beyond the Black Box*. University of Cambridge, School of Education (www.nuffieldfoundation.org/sites/default/files/files/beyond_blackbox.pdf)

Atar B (2011) An application of descriptive and explanatory item response models to TIMSS 2007 Turkey mathematical data, *Egitim Ve Bilim-Education and Science*, 36, 255–269

Atar B & Aktan D C (2013) Person Explanatory Item Response Theory Analysis: Latent Regression Two Parameter Logistic Model, *Egitim Ve Bilim-Education and Science*, 38 (168), 59–68

Ayala, C C, Shavelson, R J, Ruiz Primo, M A, Brandon, P R, Yin, Y, Furtak, E M, Young, D B & Tomita, M (2008) From formal embedded assessments to reflective lessons: The development of formative assessment studies, *Applied Measurement in Education*, 21 (4), 315

Au, W (2007) High-stakes testing and curricular control: a qualitative metasynthesis, *Educational Researcher*, 36 (5), 258–267

Azúa, X & Bick, M (2009) Formación de profesores en evaluación para el aprendizaje [Teacher education on assessment for Learning]. In: C Sotomayo & H Walker (eds.) *Formación Continua de Profesores*. Santiago de Chile, Editorial Universitaria, 263–283.

Babiar T C (2011) Exploring differential item functioning (DIF) with the Rasch model: a comparison of gender differences on eighth grade science items in the United States and Spain, *Journal of Applied Measurement*, 12 (2), 144–164

Bachman, L F (2005) Building and supporting a case for test use, *Language Assessment Quarterly*, 2 (1), 1–34

Bachmann, K, Haug, P & Myklebust, R (2010) Med rett til å prestere. In; E Elstad & K Sivesind (eds.) *I PISA – Sannheten om skolen?* Oslo, Universitetsforlaget

Bachman, L F & Palmer, A (2010) *Language Assessment in Practice*. Oxford, Oxford University Press

Baird, J & Black, P (2013) Test theories, educational priorities and reliability of public examinations in England, *Research Papers in Education*, 28 (1), 5–21

Baird, J, Isaacs, T, Johnson, S, Stobart, G, Yu, G, Sprague, T & Daugherty, R (2011) *Policy Effects of PISA*. Report commissioned by Pearson UK http://oucea.education.ox.ac.uk/wordpress/wp-content/uploads/2011/10/Policy-Effects-of-PISA-OUCEA.pdf

Baker, E L (2000) *Understanding Educational Quality: Where Validity Meets Technology*. William H. Angoff Memorial Lecture Series. Princeton, NJ, Educational Testing Service

Baker, E L (2013) The Chimera of Validity, *Teachers College Record*, 115 (9) (www.tcrecord.org/library) ID Number: 17106

Baker, E L & Linn, R L (2002) *Validity Issues for Accountability Systems*. CSE Technical Report 585. LA, California, National Center for Research on Evaluation, Standards, and Student Testing, University of California

Ball, S J (2008) *The education debate*. Bristol, UK, The Policy Press

Ball, S J, Maguire, M & Braun, A (2012) *How schools do policy. Policy enactment in secondary schools*. Abingdon, UK, Routledge

Balkanyi, P (2012) The Impact of PIRLS in Hungary. In: K Schwippert & J Lenkeit (eds.) *Progress in Reading Literacy in National and International Context.* Studies in International Comparative and Multicultural Education, Vol 13, The Impact of PIRLS 2006 in 12 countries. Munster, Waxman

Bandura, A (1971) *Social Learning Theory*. Englewood Cliffs, NJ, Prentice-Hall

Bangert-Drowns, R L, Kulik, C C, Kulik, J A & Morgan, M (1991) The instructional effect of feedback in test-like events. *Review of Educational Research*, 61 (2), 213–238

Behrens, J T, Mislevy, R J, DiCerbo, K E & Levy, R (2012) Evidence centered design for learning and assessment in the digital world. In: M C Mayrath, J Clarke-Midura & D H Robinson (eds.) *Technology-Based Assessments for 21st Century Skills. Theoretical and practical implications from modern research.* Charlotte, NC, Information Age Publishing, 13–53

Bennett, R E (2010) Cognitively Based Assessment of, for, and as Learning (CBAL): a preliminary theory of action for summative and formative assessment, *Measurement: Interdisciplinary Research and Perspectives*, 8 (2–3), 70–91

Bennett, R E (2011) Formative assessment: a critical review, *Assessment in Education: Principles, Policy & Practice*, 18 (1), 5–25

Bennett, R E, Kane, M & Bridgeman, B (2011) *Theory of Action and Validity Argument in the Context of Through-Course Summative Assessment*. Princeton, NJ, Educational Testing Service

Bergem, O K, Grønmo, L S & Olsen, R V (2005) PISA 2003 og TIMSS 2003. Hva forteller disse undersøkelsene om norske elevers kunnskaper og ferdigheter i matematikk?, *Norsk Pedagogisk Tidsskrift*, 89 (1), 31–44

Bernstein, B (1996) *Pedagogy, symbolic control and identity.* London, Taylor&Francis

Berryhill, J, Linney, J A & Fromewick, J (2009) The effects of education accountability on teachers: are policies too stress-provoking for their own good?, *International Journal of Education Policy and Leadership*, 4 (5) (www.ijepl.org)

Bernstein, D A, Clarke-Stewart, A, Roy, E J, Srull, T K & Wickens, C D (1994) *Psychology* (Third edition). Boston, Houghton Mifflin

Bew, P (2011) *Independent Review of Key Stage 2 Testing, Assessment and Accountability*. Final Report. London, Department for Education

Biggs, J (1988) Assessment and classroom learning: a role for summative assessment?, *Assessment in Education: Principles, Policy & Practice*, 5 (1), 103–110

Black, H (1986) Assessment for learning. In: D Nuttall (ed.) *Assessing educational achievement*. London, Falmer Press

Black, P (1999) Assessment, Learning Theories and Testing Systems. In: P Murphy (ed.) *Learners, Learning and Assessment*. London, Paul Chapman, 118–134

Black, P with the King's College London Assessment for Learning Group (Harrison, C, Lee, C, Marshall, B & Wiliam, D) (2004) *The Nature and Value of Formative Assessment for Learning* (Draft). London, King's College London

Black, P & Wiliam, D (1998a) Assessment and Classroom Learning, *Assessment in Education: Principles, Policy & Practice*, 5 (1), 7–74

Black, P & Wiliam, D (1998b) *Inside the Black Box: Raising Standards Through Classroom assessment*. London, King's College London School of Education.

Black, P & Wiliam, D (2009) Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21 (5), 5–31

Black, P, Harrison, C, Lee, C, Marshall, B & Wiliam, D (2003) *The nature and value of formative assessment for learning*. Paper presented at the Annual Meeting of American Educational Research Association, Chicago, April

Black, P, Harrison, C, Lee, C, Marshall, B & Wiliam, D (2010) *Assessment for Learning, Putting it into practice*. London, Open University Press

Blinkhorn, S F (1997) Past Imperfect, Future Conditional: Fifty Years of Test Theory, *British Journal of Mathematical and Statistical Psychology*, 50 (2), 175–185

Bloom, B S (1971) Mastery Learning. In: J Block (ed.) *Mastery Learning: Theory and Practice*. New York, Holt, Rinehart & Winston, 47–63

Bloom, B S, Hastings, J T & Madaus, G F (eds.) (1971) *Handbook of formative and summative evaluation of student learning*. New York, McGraw-Hill

Bloom, B S (ed.) (1956) *Taxonomy of Educational Objectives*. New York, David McKay Co. Inc

Bloom, B S (1986) *Learning for mastery.* Evaluation Comment. (UCLA-CSIEP), 1 (2), 1–2

Board of Education (1911) *Report of the Consultative Committee on Examinations in Secondary Schools*. London, His Majesty's Stationery Office

Boden, M A (1988) *Computer Models of Mind*. Cambridge, Cambridge University Press

Bonnet G (2002) Reflections in a critical eye: on the pitfalls of international assessment, *Assessment in Education: Principles, Policy & Practice*, 9 (3), 387–399

Borsboom, D (2005) *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Cambridge, Cambridge University Press

Borsboom, D (2006) The Attack of the Psychometricians, *Psychometrika*, 71 (3), 425–440

Borsboom, D & Mellenbergh, G J (2007) Test validity in cognitive assessment. In: J P Leighton & M J Gierl (eds.) *Cognitive Diagnostic Assessment for Education: Theory and Applications*. New York, Cambridge University Press, 85–115

Borsboom, D, Cramer, A O J, Keivit, R A, Scholten, A Z & Franic, S (2009) The end of construct validity. In: R W Lissitz (ed.) *The Concept of Validity: Revisions, New Directions, and Applications*. Charlotte, NC, Information Age Publishing, 135–170

Borsboom, D, Mellenbergh, G J & van Heerden, J (2004) The concept of validity, *Psychological Review*, 111 (4), 1061–1071

Bracey G W (2000) The TIMSS 'Final Year' Study and report: a critique, *Educational Researcher*, 29 (4), 4–10

Brandon, P R (2004) Conclusions About Frequently Studied Modified Angoff Standard-Setting Topics, *Applied Measurement in Education*, 17 (1), 59–88

Brandsford, J D, Brown, A L & Cocking, R R (2000) *How People Learn. Brain, Mind, Experience, and School*. Washington National Academy Press

Breakspear, S (2012) *The Policy Impact of PISA: An Exploration of the Normative Effects of International Benchmarking in School System Performance*. OECD Education Working Papers, No. 71. OECD Publishing

Bredo, E (1997) The social construction of learning. In: G D Phye (ed.) *Handbook of Academic Learning.* San Diego CA, Academic Learning, 2–45

Brennan, R L (2001) (Mis)conceptions about generalizability theory, *Educational Measurement: Issues and Practice*, Spring, 5–10

Brennan, R L (2013) Commentary on 'Validating the Interpretations and Uses of Test Scores', *Journal of Educational Measurement*, 50 (1), 74–83

Briggs, D C, Ruiz-Primo, M A, Furtak, E, Shephard, L & Y. Yin (2012) Meta-Analytica Methodology and Inferences about the Efficacy of Formative Assessment, *Educational Measurement: Issues and Practice*, 31, 13–17

Briggs, D C, Alonzo, A C, Schwab, C & Wilson M (2006) Diagnostic assessment with ordered multiple-choice items, *Educational Assessment*, 11 (1), 33–63

Brookhart, S M (2001) Successful students' formative and summative uses of assessment information, *Assessment in Education: Principles, Policy & Practice*, 8 (2), 153–169

Brookhart, S M (2004) Assessment theory for college classrooms*, New Directions for Teaching and Learning,* 5–14

Brookhart, S M (2004) Classroom assessment: Tensions and intersections in theory and practice, *Teachers College Record*, 106 (3), 429–458

Brookhart, S M (2007) Expanding views about formative classroom assessment: A review of the literature. In: J H McMillan (ed.) *Formative classroom assessment: Theory into practice*. New York, NY, Teachers College Press, 43–62

Brown, G T L, Kennedy, K J, Fok, P K, Chan, J & Yu, W M (2009) Assessment for Student Improvement: Understanding Hong Kong Teachers' Conceptions and Practices of Assessment, *Assessment in Education: Principles, Policy & Practice*, 16 (3), 347–363

Brown, P, Lauder, H & Ashton, D (2011) *The Global Auction: the broken promises of education, jobs and income*. Oxford, Oxford University Press

Bruner, J (1960) *The Process of Education*. New York, Vintage Books

Brunner, M, Artfelt, C, Krauss, S & Baumert, J (2007) Coaching for the PISA test, *Learning and Instruction*, 17 (2), 111–122

Bulle, N (2011) Comparing OECD educational models through the prism of PISA, *Comparative Education*, 47 (4), 503–521

Carless, D (2005) Prospects for the Implementation of Assessment for Learning. *Assessment in Education Principles Policy & Practice*, 12 (1), 39–54

Carless, D (2011) From Testing to Productive Student Learning: Implementing Formative Assessment in Confucian-Heritage Settings, *Journal of Second Language Teaching and Research,* 52, 25–45

Cao, J & Stokes S L (2008) Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika*, 73 (2), 209–230

Caro, D H & Lenkeit, J (2012) An analytical approach to study educational inequalities: 10 hypothesis tests in PIRLS 2006, *International Journal of Research in Method & Education*, 35 (1), 3–30

Caro, D H, Sandoval-Hernandez, A & Lüdtke, O (2013) Cultural, Social and Economic Capital Constructs in International Assessments: An Evaluation Using Exploratory Structural Equation Modeling, School Effectiveness and School Improvement, *School Effectiveness and School Improvement*, published online

Carvalho, L M & Costa, E (2009) Production of OECD's 'Programme for International Student Assessment': final report. Project KNOWandPOL, WP 11, March http://knowandpol.eu/IMG/pdf/o31.pisa.portugal.pdf

Chapelle, C A, Enright, M K & Jamieson, J (2010) Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, 29 (1), 3–13

Chapelle, C A, Enright, M K & Jamieson, J (eds.) (2008) *Building a Validity Argument for the Test of English as a Foreign Language*. London, Routledge

Cherryholmes, C H (1988) Construct validity and the discourses of research, *American Journal of Education*, 96, 421–457

Cizek, G. (2010) An Introduction to Formative Assessment: History, Characteristics, and Challenges. In: H L Andrade & G J Cizek (eds.) *Handbook of Formative Assessment*. New York, Routledge

Cizek, G J, Bunch, M B & Koons, H (2004) Setting performance standards: contemporary methods. An NCME Instructional Module, *Educational Measurement: Issues and Practice*, 23 (4), 31–50

Cizek, G J (2012) Defining and distinguishing validity: interpretations of score meaning and justification of test use, *Psychological Methods,* 17 (1), 31–43

Cliff, N (1992) Abstract measurement theory and the revolution that never happened, *Psychological Science*, 3 (3), 186–190

Cobb, P (1999) Where is the mind? In: P Murphy (ed.) *Learners, Learning and Assessment*. The Open University, Paul Chapman Publishing, 135–150

Coffey, J E, Hammer, D, Levin, D M & Grant, T (2011) The missing disciplinary substance of formative assessment, *Journal of Research in Science Teaching*, 48 (10), 1109–1136

Colwill, I (2007) *Improving GCSE: Internal and Controlled Assessment.* London, Qualifications and Curriculum Authority

Condie, R, Livingston, K & Seagraves, L (2005) *Evaluation of the Assessment is for Learning Programme: final report*. Glasgow, Quality in Education Centre, University of Strathclyde

Crocker, L (1997) Editorial: The great validity debate, *Educational Measurement: Issues and Practice*, 16 (2), 4

Crooks, T J (1988) The impact of classroom evaluation practices on students, *Review of Educational Research*, 58 (4), 438–481

Cronbach, L J (1949) Essentials of Psychological Testing. New York, Harper & Brothers

Cronbach, L J (1971) Test validation. In: R L Thorndike (ed.) *Educational Measurement* (Second edition). Washington, DC, American Council on Education, 443–507

Cronbach, L J (1980a) Validity on parole: How can we go straight? In: W B Schrader (ed.) *Measuring Achievement: Progress Over a Decade*. Proceedings of the 1979 ETS Invitational Conference San Francisco, CA, Jossey-Bass, 99–108

Cronbach, L J (1980b) Selection theory for a political world, *Public Personnel Management*, 9 (1), 37–50

Cronbach, L J (1988) Five perspectives on validity argument. In: H Wainer & H I Braun (eds.) *Test Validity*. Hillsdale, NJ, Lawrence Erlbaum, 3–17

Cronbach, L J (1964) *Essentials of Psychological Testing*. Second Edition. New York, Harper & Row

Crooks, T J (1988) The impact of classroom evaluation practices on students, *Review of Educational Research*, 58 (4), 438–481

Crooks, T J (2011) Assessment for learning in the accountability era: New Zealand, *Studies in Educational Evaluation*, 37 (1), 71–77

Crossouard, B (2011) Using formative assessment to support complex learning in conditions of social adversity, *Assessment in Education: Principles, Policy & Practice*, 18 (1), 59–72

Cureton, E E (1951) Validity. In: E F Lindquist (ed.) *Educational Measurement*. Washington, DC, American Council on Education, 621–694

Daly, A L, Baird, J, Chamberlain, S & Meadows, M (2012) Assessment reform: students' and teachers' responses to the introduction of stretch and challenge at A-level, *The Curriculum Journal*, 23 (2), 139–155

de Corte, E (2010) Historical developments in the understanding of learning. In: H Dumont, D Istance & F Benavides (eds.) *The Nature of Learning. Using research to inspire practice*. OECD (www.educ.ethz.ch/pro/litll/oecdbuch.pdf)

DeLuca, C (2011) Interpretive validity theory: mapping a methodology for validating educational assessments, *Educational Research*, 53 (3), 303–320

Dempster, F N (1991) Synthesis of research on reviews and tests, *Educational Leadership*, 48 (7), 71–76

Dempster, F N (1992) Using tests to promote learning: A neglected classroom resource, *Journal of Research and Development in Education*, 25 (4), 213–217

Dempster E R & Reddy V (2007) Item readability and science achievement in TIMSS 2003 in South Africa, *Science Education*, 91 (6), 906–925

Dewey, J (1897) My Pedagogic Creed, *School Journal*, 54, 77–80
Dewey, J (1916) *Democracy and Education*. New York, Macmillan
DfES (Department for Education and Skills) (2007) *Assessment for learning 8 schools project report*. London, DfES Publications, London
Darling-Hammond, L & Rustique-Forrester, E (2005) The consequences of student testing for teaching and teacher quality. In: J L Herman & E H Haertel (eds.) *Uses and Misuses of Data for Educational Accountability and Improvement*. The 104th Yearbook of the National Society for the Study of Education, Part 2. Malden, MA, Blackwell Publishing, 289–319
Dominguez, M, Viera, M J & Vidal, J (2012) The impact of the Programme for International Student Assessment on academic journals, *Assessment in Education: Principles, Policy & Practice*, 19 (4), 393–409
Dori, Y J (2003) From nationwide standardized testing to school-based alternative embedded assessment in Israel: Students' performance in the Matriculation 2000 project, *Journal of Research in Science Teaching*, 40 (1), 34–52
Dumont, H, Istance, D & Benavides, F (eds.) (2010) *The Nature of Learning. Using research to inspire practice*. Organisation for Economic Co-operation and Development (www.educ.ethz.ch/pro/litll/oecdbuch.pdf)
Dunn, K & Mulvenon, S (2009) A Critical Review of Research on Formative Assessment: The Limited Scientific Evidence of the Impact of Formative, *Practical Assessment, Research and Evaluation*, 14 (7), 1–11
Dysthe, O (2008) Klasseromsvurdering og læring, *Bedre skole*, 3, 16–23
Ecclestone, K (2002) *Learning autonomy in post-16 education: policy and practice in formative assessment.* London, Falmer-Routledge
Egelund, N (2008) The value of international comparative studies of achievement: a Danish perspective. *Assessment in Education: Principles, Policy & Practice*, 15 (3), 245–251
Eggen, A E (2007) Vurderingskompetanse og definisjonsmakt, *Norsk Pedagogisk Tidsskrift*, 2, 150–163
Eklöf, H (2010) Skill and will: test-taking motivation and assessment quality, *Assessment in Education: Principles, Policy & Practice*, 17 (4), 345–356
Elshout-Mohr, M (1994) Feedback in self-instruction, *European Education*, 26 (2), 58–73
Elstad, E, Nortvedt, G A & Turmo, A (2009) The Norwegian Assessment System: An Accountability Perspective, *CADMO* 17 (1), 89–103
Elstad, E (2012) PISA debates and blame management among the Norwegian educational authorities: press coverage and debate intensity in the newspapers, *Problems of Education in the 21st Century*, 48, 10–22
Elvers E (2010) PISA: Issues in implementation and interpretation, *The Irish Journal of Education*, 38, 94–118
Elwood, J (2006) Gender Issues in Testing and Assessment. In: C Skelton, B Francis & L Smulyan (eds.) *The SAGE Handbook on Gender and Education*. SAGE, 262–278
Elwood, J & Klenowski, V (2002) Creating communities of shared practice: the challenges of assessment use in learning and teaching, *Assessment and Evaluation in Higher Education*, 27 (3), 243–256
Embretson, S (1993) Psychometric models for learning and cognitive processes. In: N Fredriksen, R J Mislevy & I I Bejar (eds.) *Test theory for a new generation of tests*. Psychology Press, 125–150
Ercikan, K & Koh, K (2009) Examining the construct comparability of the English and French version of TIMSS. *International Journal of Testing*, 5 (1), 23–35
Engestrom, Y (1987) *Learning by expanding: an activity theoretical approach to developmental research.* Helsinki, Orienta-Konsultit
Ertl, H (2006) Educational standards and the changing discourse on education: the reception and consequences of the PISA study in Germany, *Oxford Review of Education*, 32 (5), 619–634
Forer, B & Zumbo, B D (2011) Validation of multilevel construct: Validation methods and empirical findings for the EDI, *Social Indicators Research*, 103, 231–265

Forte Fast, E & Hebbler, S with ASR-CAS Joint Study Group on Validity in Accountability Systems (2004) *A Framework for Examining Validity in State Accountability Systems*. Washington, DC, Council of Chief State School Officers

Frederiksen, J R & Collins, A (1989) A systems approach to educational testing, *Educational Researcher*, 18 (9), 27–32

Fremer, J (2000) Promoting high standards and the 'problem' with construct validation, *NCME Newsletter*, 8 (3), 1

Frey A & Seitz N N (2011) Hypothetical Use of Multidimensional Adaptive Testing for the Assessment of Student Achievement in the Programme for International Student Assessment, *Educational and Psychological Measurement*, 71 (3), 503–522

Fuchs, L S & Fuchs, D (1986) Effects of systematic formative evaluation – A metaanalysis, *Exceptional Children*, 53 (3), 199–208

Frønes, T S, Roe, A & Vagle, W (2012) Nasjonale prøver i lesing – utvikling, resultater og bruk. In: T N Hopfenbeck, M Kjærnsli & R V Olsen (eds.) *Kvalitet i norsk skole: Internasjonale og nasjonale undersøkelser av læringsutbytte og undervisning*. Oslo, Universitetsforlaget, 135–153

Frønes, T, Narvhus, E & Jetne, Ø (2011) *Elever på nett - Digital lesing i PISA 2009*. UiO, Institutt for lærerutdanning og skoleforskning

Gagné, R M (1985) *The Conditions of Learning and Theory of Instruction (4th Edition)*. New York, CBS College Publishing

Gardner, J (ed.) (2012) *Assessment and Learning* (Second edition). London, SAGE Publications Ltd

Gardner, J, Harlen, W, Hayward, L & Stobart, G (2011) Engaging and empowering teachers in innovative assessment practice. In: R Berry & B Adamson (eds.) *Assessment Reform in Education: Policy and Practice*. London, Springer, 105–119

Gergen, K J & Dixon-Román, E J (2013) *Epistemology and measurement: paradigms and practice. II Social epistemology and the pragmatics of assessment*. The Gordon Commission on the Future of Assessment in Education (www.gordoncommission.org/rsc/pdf/dixonroman_gergen_epistemology_measurement_paradigms_practices_2.pdf)

Gioka, O (2006) Assessment for Learning in Physics Investigations: Assessment Criteria, Questions and Feedback in Marking, *Physics Education*, 41 (4), 342–346

Gipps, C (1994) *Beyond Testing: Towards a Theory of Educational Assessment*. London, Falmer

Gipps, C, McCallum, B, Hargreaves, E & Pickering, A (2005) *From TA to Assessment for Learning: the impact of assessment policy on teachers' assessment practice*. Paper presented at the British Educational Research Association Annual Conference, University of Glamorgan, 14–17 September 2005

Goldstein, H (1979) Consequences of Using the Rasch Model for Educational Assessment, *British Educational Research Journal,* 5 (2), 211–220

Goldstein, H (1980) Dimensionality, Bias, Independence and Measurement Scale Problems in Latent Trait Test Score Models, *British Journal of Mathematical and Statistical Psychology,* 33 (2), 234–246

Goldstein, H (2004) International comparisons of student attainment: some issues arising from the PISA study, *Assessment in Education: Principles, Policy & Practice*, 11 (3), 319–330

Goldstein, H (2012) Francis Galton, measurement, psychometrics and social progress, *Assessment in Education: Principles, Policy & Practice*, 19 (2), 147–158

Goldstein, H, Bonnet G & Rocher T (2007) Multilevel structural equation models for the analysis of comparative data on educational performance*, Journal of Educational and Behavioral Statistics*, 32 (3), 252–286

Good, F J & Cresswell, M J (1988) Grade awarding judgments in differentiated examinations, *British Educational Research Journal*, 14, 263–281

Greger, D (2012) When PISA does not matter? The case of the Czech Republic and Germany, *Human Affairs*, 22 (1), 31–42

Grek, S (2009) Governing by the numbers: the PISA 'effect' in Europe, *Journal of Education Policy*, 24 (1), 23–27

Grek, S (2010) International Organisations and the Shared Construction of Policy 'Problems': problematisation and change in education governance in Europe, *European Educational Research Journal*, 9 (3), 396–406

Grisay, A (2003) Translation procedures in OECD/PISA 2000 international assessment, *Language Testing*, 20, 225–240

Grisay, A, de Jong J H, Gebhardt E, Berenzer A & Halleux-Monseur B (2007) Translation equivalence across PISA countries, *Journal of Applied Measurement*, 8 (3), 249–266

Grisay, A & Gonzalez E & Monseur C (2009) Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments, *IERI Monograph Series: Isues and methodologies in large-scale assessments*, 63–83

Grisay, A & Monseur C (2007) Measuring the equivalence of item difficulty in the various versions of an international test, *Studies in Educational Evaluation*, 33, 69–86

Grønmo, L S, Bergem, O K, Kjærnsli, M, Lie, S & Turmo, A (2004) *Hva I all verden har skjedd i realfagene? Norske elevers prestasjoner I matematikk og naturfag I TIMSS 2003*. Oslo, Universitetet i Oslo

Grønmo, L S & Onstad, T (2009) *TIMSS 2007: Tegn til bedring?* Oslo, Unipub

Gulikers, J T M, Bastiaens, T J & Kirschner, P A (2004) A five-dimensional framework for authentic assessment, *Educational Technology Research and Development*, 52 (3), 67–86

Gunnes, H, Rorstad, K 7 D W Aksnes (2013) *Utdanningsforskning i Norge 2011: Ressurser og resultater*. NIFU rapport 31

Gür, B S, Çelik, Z & Özoglu, M (2012) Policy options for Turkey: a critique of the interpretation and utilization of PISA results in Turkey, *Journal of Education Policy*, 27 (1), 1–21

Guskey, T R (2010) Formative Assessment: The Contribution of Benjamin S. Bloom. In: H L Andrade & G J Cizek (eds.) *Handbook of Formative Assessment*. Abingdon, UK, Routledge

Gustafsson J-E (2012) Något om utveklingen av de internationallea studerna av kunskaper og ferdigheter. In: T N Hopfenbeck, M Kjærnsli & R V Olsen (eds.) *Kvalitet i norsk skole. Internasjonale og nasjonale undersøkelser av læringsutbytte og undervisning*. Oslo, Universitetsforlaget

Gustafsson J-E & Rosen G (2006) The dimensional structure of reading assessment tasks in the IEA reading literacy study 1991 and the Progress in International Reading Literacy Study 2001, *Educational Research and Evaluation: An International Journal on Theory and Practice*, 12 (5), 445–468

Gutiérrez, K D & Penuel, W R (2014) Relevance to practice as a criterion for rigor, *Educational Researcher*, 43 (1), 19–24

Haertel, E H (2013) How is testing supposed to improve schooling?, *Measurement: Interdisciplinary Research and Perspectives*, 11 (1–2), 1–18

Haertel, E H & Herman, J L (2005) A historical perspective on validity arguments for accountability testing. In: J L Herman & E H Haertel (eds.) *Uses and Misuses of Data for Educational Accountability and Improvement*. The 104th Yearbook of the National Society for the Study of Education, Part 2. Malden, MA, Blackwell Publishing, 1–29

Haladyna, T M & Rodriguez M C (2013) *Developing and Validating Test Items*. Routledge, New York

Hamilton, L S, Stecher, B M & Yuan, K (2012) Standards-based accountability in the United States: lessons learned and future directions, *Education Inquiry*, 3 (2), 149–170

Hammersley, M (2001) On 'systematic' reviews of research literature: a narrative reply to Evans and Benefield, *British Educational Research Journal*, 27 (5), 543–554

Hamp-Lyons, L (1997) Washback, impact and validity: Ethical concerns. *Language Testing*, 14 (3), 295–303

Hanson, F A (1993) *Testing Testing. Social consequences of the examined life*. London, University of California Press

Hanson, F A (2000) How tests create what they are intended to measure. In Filer, A (ed.) *Assessment. Social Practice and Social Product*. London, Routledge, 67–81

Harrison, C (2013) Collaborative action research as a tool for generating formative feedback on teachers' classroom assessment practice: the KREST project, *Teachers and Teaching,* 19 (2), 209–220

Harlen, W (2007) Criteria for evaluating systems for assessment, *Studies in Educational Evaluation*, 33, 15–28

Harlen, W & Deakin Crick, R (2002) *A systematic review of the impact of summative assessment and tests on students' motivation for learning*. London, EPPI-Centre, Social Science Research Unit, Institute of Education

Harlen, W & James, M (1999) Assessment and Learning: differences and relationshops between formative and summative assessment, *Assessment in Education*: *Principles, Policy & Practice*, 4 (3), 365–379

Hattie, J & Timperley, H (2007) The power of feedback, *Review of Educational Research*, 77 (1), 81–112

Hattie, J (2009) *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. London, Routledge

Haug, P (2003) The Research Council of Norway Evaluation Reform 97. In: P Haug & T A Schwandt (eds.) *Evaluating Educational Reforms: Scandinavian Perspectives*. Greenwich, CT, Information Age Publishing

Haug, P (2004a) Sentrale resultat fra evalueringa av Reform 97, *Norsk Pedagogisk Tidsskrift*, 88 (4), 284–263

Haug, P (2004b) *Resultat frå evalueringa av Reform 94.* Oslo, Norges forskningsråd

Haug, P & Monsen, L (eds.) (2002) *Skolebasert vurdering – erfaringer og utfordringer.* (School-based evaluation -experience and challenges). Oslo, Abstrakt forlag

Hauger J & Sireci S (2008) Detecting Differential Item Functioning Across Examinees Tested in Their Dominant Language and Examinees Tested in a Second Language, *International Journal of Testing*, 8 (3), 237–250

Hayward, L & Spencer, E (2010) The Complexities of Change: Formative Assessment in Scotland, *Curriculum Journal*, 21 (2), 161–177

Hertzberg, F (2003) Arbeid med muntlige ferdigheter. In: K. Klette (ed) Klasserommets praksisformer etter reform 97 (Vol. 1/03, s 226) Oslo, Unipub

Hertzberg, F (2008) Assessment of writing in Norway – a case of balancing dilemmas. In: A Havnes & L McDovell (eds.) *Balancing Dilemmas in Assessment and Learning in Contemporary Education*. New York, Routledge, 51–60

Higham, J & Yeomans, D (2010) Working together? Partnership approaches to 14-19 education in England, *British Educational Research Journal*, 36 (3), 379–401

Hilton, M (2006) Measuring standards in primary English: issues of validity and accountability with respect to PIRLS and National Curriculum test scores, *British Educational Research Journal*, 32 (6), 817–37

Hilton, M (2007) A further brief response from Mary Hilton to 'Measuring standards in primary english: the validity of PIRLS – a response to Mary Hilton' by Chris Whetton, Liz Twist and Marian Sainsbury, *British Educational Research Journal*, 33 (6), 987–990

Hilborne, N (2005) Coursework is a 'charter to cheat', *The Times Educational Supplement*, 25 November (www.tes.co.uk/teaching-resource/Coursework-is-a-charter-to-cheat-2169722)

Holzinger, K & Knill, C (2005) Causes and conditions of cross-national policy convergence, *Journal of European Public Policy*, 12, 775–796

Hollingworth, L (2012) Why leadership matters: empowering teachers to implement formative assessment, *Journal of Educational Administration* 50 (3), 365–379

Hopfenbeck, T N (2011) Fostering Self-regulated Learners in a Community of Quality Assessment Practices, *CADMO*, (1) 7–21

Hopfenbeck, T N, Tolo, A, Florez, M T & El Masri, Y (2013) *Balancing Accountability and Trust*. Paris, Organisation for Economic Co-operation and Development

Hounsell, D et al. (2008) The quality of guidance and feedback to students, *Higher Education Research & Development* 27, (1), 55-67

House of Commons Children, Schools and Families Committee (2008) *Testing and Assessment. Third Report of Session 2007–08. Volume I*. London, The Stationery Office Limited

Hout, M & Elliott, S W (eds.) (2011) *Incentives and Test-Based Accountability in Education*. Committee on Incentives and Test-Based Accountability in Public Education of the National Research Council. Washington, DC, National Academies Press

Hutchinson, D, Kendall, L, Bartholomew, D, Knott, M, Galbraith, J & Piccoli, M (2000) Reliability of assessment of reading ability in three countries, *Quality & Quantity*, 34, 353–365

Ibrahim, R (2009) Psycholinguistic challenges in processing Arabic languages, *Second Languges: Teaching, learning and assessment*, 61–82

Impara, J C & Plake, B S (1998) Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method, *Journal of Educational Measurement*, 35, 69–81

Imsen, G (2011) Hattie-feberen i norsk skolepolitikk (The Hattie-fever in Norwegian School policy). *Bedre skole*, 4

Imsen, G (2004) *Democratic evaluation and evaluation for learning. A reply to Ove Karlsson*. In: P Haug & T Schwandt (eds.) *Evaluating Educational Reforms Scandinaivan Perspectives.* A volume in Evaluation and Society, Connecticut, IAP Information Age Publishing

James, M (2006) Assessment and learning. In: S Swaffield (ed.) *Unlocking Assessment. Understanding for reflection and application*. Abingdon, UK, Routledge, 20–35

James, M (1992) *Assessment for Learning*. Paper presented at the Annual Conference of the Association for Supervision and Curriculum Development, April. New Orleans, LA

James, M., McCormick, R., Black, P., Carmichael, P., Drummond, M. -J., & Fox, A., et al. (2007). *Improving learning how to learn: Classrooms, schools and networks*. London: Routledge.

James, M (2012) Growing confidence in educational research: threats and opportunities, *British Educational Research Journal*, 38 (2), 181–201

James, M & Lewis, J (2012) Assessment in harmony with our understanding of learning: problems and possibilities. In: J Gardner (ed.) *Assessment and Learning* (Second edition). London, SAGE, 187–205

James, W (1890) *The Principles of Psychology*. New York, Henry Holt and Company

James, W (1899, 1958) *Talks to teachers on psychology; And to students on some of life's ideals.* New York, W W Norton and Co

Jesson, J, Matheson, L & Lacey, F M (2011) *Doing your literature review. Traditional and systematic approaches*. London, SAGE

Kane, M T (1992) An argument-based approach to validity, *Psychological Bulletin*, 112 (3), 527–535

Kane, M T (2001) Current concerns in validity theory, *Journal of Educational Measurement*, 38 (4), 319–342

Kane, M T (2002) Validating high-stakes testing programs, *Educational Measurement: Issues and Practice*, 21 (1), 31–41

Kane, M T (2004) Certification testing as an illustration of argument-based validation, *Measurement: Interdisciplinary Research and Perspectives*, 2 (3), 135–170

Kane, M T (2006) Validation. In: R L Brennan (ed.) *Educational Measurement* (Fourth edition). Washington, DC, American Council on Education/Praeger, 17–64

Kane, M T (2009) Validating the interpretations and uses of test scores. In: R W Lissitz (ed.) *The Concept of Validity: Revisions, New Directions, and Applications*. Charlotte, NC, Information Age Publishing, 39–64

Kane, M T (2013) Validating the interpretations and uses of test scores, *Journal of Educational Measurement*, 50 (1), 1–73

Kellard, K, Costello, M, Godfrey, D, Griffiths, E & Rees, C (2008) *Evaluation of the Developing Thinking and Assessment for Learning development programme*. Welsh Assembly Government

Kennedy, J K, Chan, J K S, Fok, P K, & Yu, W M (2008) Forms of assessment and their potential for enhancing learning: conceptual and cultural issues, *Educational Research Policy and Practice*, 7, 197-207

Kingston, N & Nash, B (2011) Formative Assessment: A Meta-Analysis and a Call for Research, *Educational Measurement: Issues and Practice*, 30, 28–37

Kirton, A, Hallam, S, Peffers, J, Robertson, P & Stobart, G (2007) Revolution, evolution or a Trojan horse? Piloting assessment for learning in some Scottish primary schools, *British Educational Research Journal*, 33 (4), 605–627

Kjærnsli, M, Lie, S, Olsen, R V, Roe, A & Turmo, A (2004) *Rett spor eller ville veier? Norske elevers prestasjoner I matematikk, naturfag og lesing I PISA 2003*. Oslo, Universitetsforlaget

Kjærnsli, M, Lie, S, Olsen, R V & Roe, A, (2007) *Tid for tunge loft, Norske Elevers Kompetanse I Naturfag, Lesing og Matematikk i PISA 2006*. Oslo, Universietetsforlaget

Klenowski, V (2009) Assessment for Learning revisited: an Asia-Pacific perspective, *Assessment in Education: Principles, Policy & Practice,* 16 (3), 263–68

Klette, K (ed.) (2003) *Livet i klasserommet: Arbeids- og samtaleformer* (Vol. 1/03). Oslo, Unipub

Klette, K & Lie, S (2006) *Sentrale funn. Forelopige resulater fra PISA+ prosjektet*. Oslo, Universitetet i Oslo, Det utdanningsvitenskapelige fakultet

Kluger, A N & DeNisi, A (1996) The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory, *Psychological Bulletin*, 119 (2), 254–284

Koch, M J (2013) The multiple-use of accountability assessments: implications for the process of validation, *Educational Measurement: Issues and Practice*, 32 (4), 2–15

Koch, M J & DeLuca, C (2012) Rethinking validation in complex high-stakes assessment contexts, *Assessment in Education: Principles, Policy & Practice*, 19 (1), 99–116

Köller, O (2005) Formative assessment in classrooms: A review of the empirical German literature. In: J Looney (ed.) *Formative assessment: Improving learning in secondary classrooms*. Paris, Organisation for Economic Co-operation and Development, 265–279

Koretz, D (2008) *Measuring Up: What Educational Testing Really Tells Us.* Cambridge, MA, Harvard University Press

Koretz, D (2013) Commentary on Haertel, E, 'How is testing supposed to improve schooling?' *Measurement: Interdisciplinary Research and Perspectives*, 11 (1–2), 40–43

Kreiner, S & Christensen, K B (2013) Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy, *Psychometrika*, published online

Kellaghan, T (2004) *Assessing student learning in Africa*. Washington, DC, World Bank

Kyvik, S & Olsen (2013) Increasing completion rates in Norwegian doctoral training: multiple causes for efficiency improvements, *Studies in Higher Education*, 38, 1–15

Kyvik, S (2014) Assessment procedures of Norwegian PhD theses as viewed by examiners from the USA, the UK and Sweden, *Assessment & Evaluation in Higher Education*, 39 (2), 140–153

Lane, S, Parke, C S & Stone, C A (1998) A framework for evaluating the consequences of assessment programs, *Educational Measurement: Issues and Practice*, 17 (2), 24–28

Lauder, H, Dillabough, J & Halsey, A, (2006) Introduction: The prospects for education: individualization, globalisation and social change. In: H Lauder, J Dillabough & A H Halsey (eds.) *Education, Globalization and Social Change*. Oxford, Oxford University Press

Lave, J & Wenger, E (1991) *Situated Learning: Legitimate Peripheral Participation*, Cambridge, Cambridge University Press.

Lawn, M (2008) *An Atlantic crossing? The work of the international examination inquiry, its researchers, methods and influence*. Symposium Books

Lawn, M & Grek, S (2012) *Europeanizing Education: Governing a new policy space*. Symposium Books

Le, L T (2009) Investigating Gender Differential Item Functioning Across Countries and Test Languages for PISA Science Items, *International Journal of Testing*, 9 (2), 122–133

Lees-Haley, P R (1996) Alice in Validityland, or the dangerous consequences of consequential validity, *American Psychologist*, 51 (9), 981–983

Lenkeit, J, Goy, M & Schwippert, K (2012) The Impact of PIRLS in Germany. In: K Schwippert & J Lenkeit, (eds.) *Progress in Reading Literacy in National and International Context*. Studies in International

Comparative and Multicultural Education, Vol. 13, The Impact of PIRLS 2006 in 12 countries. Munster, Waxman

LeTendre, G K (1999) *Competitor or Ally? Japan´s Role in American Educational Debates*. New York, Routledge

Lewy, A (1996) Postmodernism in the field of achievement testing, *Studies in Educational Evaluation*, 22 (3), 223–244

Lie, S, Kjærnsli, M, Roe, A & Turmo, A (2001) *Godt rustet for framtida? Norske 15-åringers kompetanse i lesing og realfag i et internasjonalt perspektiv.* Oslo, Institutt for lærerutdanning og skoleutvikling

Lingard, B & Grek, S (2005) *The OECD, indicators and PISA: an exploration of events and theoretical perspectives.* ESRC/ESF Research Project on Fabricating Quality in Education. Working Paper 2 (www.ces.ed.ac.uk/PDF%20Files/FabQ_WP2.pdf)

Lingard, B & Rawolle, S (2011) New scalar politics: implications for education policy,. *Comparative Education*, 47 (4), 489–502

Linn, R L (1997) Evaluating the validity of assessments: The consequences of use, *Educational Measurement: Issues and Practice*, 16 (2), 14–16

Lissitz, R W & Samuelsen, K (2007) A suggested change in terminology and emphasis regarding validity and education, *Educational Researcher*, 36 (8), 437–448

Looney, J (ed.) *Formative assessment: Improving learning in secondary classrooms.* Paris, France, Organisation for Economic Co-operation and Development, 265–279

Lorenz, E N (1972) *Predictability: Does the flap of a butterfly's wings in Brazil set off a tornado in Texas?* Presented at the 139th Meeting of the American Association for the Advancement of Science, 29 December. Sheraton Park Hotel, Washington, DC

Lord, F M & Novick, M R (1968) *Statistical theories of mental test scores*. Oxford, Addison-Wesley

Loveless, T (2014) *The 2014 Brown Center Report on American Education: How well are American students learning?* With sections on the PISA-Shanghai Controversy, Homework and the Common Core. March. Vol 3, No 3 (www.brookings.edu/research/reports/2014/03/18-brown-center-report-loveless)

Luce R D & Tukey J W (1964) Simultaneous conjoint measurement: a new type of fundamental measurement, *Journal of Mathematical Psychology*, 1, 1–27

Lyotard J-F (1984) *The Postmodern Condition: A report on knowledge (Theory and History of Literature),* Volume 10. Manchester University Press.

Maclure, M (2005) 'Clarity bordering on stupidity': where's the quality in systematic review?, *Journal of Education Policy*, 20 (4), 393–416

Macnab, D S (2000) Raising standards in mathematics education: Values, vision, and TIMSS, *Educational Studies in Mathematics,* 42, 61-80

MacPhail, A & Halbert, J (2010) 'We Had to Do Intelligent Thinking during Recent PE': Students' and Teachers' Experiences of Assessment for Learning in Post-Primary Physical Education, *Assessment in Education: Principles, Policy & Practice*, 17 (1), 23–39

Madaus, G, Russell, M & Higgins, J (2009) *The Paradoxes of High Stakes Testing. How they affect students, their parents, teachers, principals, schools and society.* Charlotte, NC, Information Age Publishing

Maguire, T, Hattie, J & Haig, B (1994) Construct validity and achievement assessment, *The Alberta Journal of Educational Research*, XL (2), 109–126

Mansell, W (2007) *Education By Numbers: The Tyranny of Testing*. London, Politico's Publishing

Maraun, M D (1998) Measurement as a normative practice: Implications of Wittgenstein's philosophy for measurement in psychology, *Theory and Psychology*, 8 (4), 435–461

Marion, S F & Pellegrino, J (2009) *Validity framework for evaluating the technical quality of alternate assessments based on alternate achievement standards*. NCME Invited Presentation. National Council on Measurement in Education Annual Meeting, 14–16 April, San Diego, California

Markus, K A & Borsboom, D (2013) *Frontiers of Test Validity Theory: Measurement, causation and meaning.* New York, Routledge

Marshall, B & Drummond, M J (2006) How teachers engage with Assessment for Learning: lessons from the classroom, *Research Papers in Education*, 21, (2), 133–149

McDowell, L et al. (2008) Evaluating assessment strategies through collaborative evidence-based practice: can one tool fit all?, *Innovations in Education and Teaching International,* 45 (2,: 143–153

McNamara, T (2001) Language assessment as social practice: challenges for research, *Language Testing*, 18 (4), 333–349

McNamara, T (2006) Validity in language testing: The challenge of Sam Messick's Legacy, *Language Assessment Quarterly*, 3 (1), 31–51

Mehrens, W A (1997) The consequences of consequential validity, *Educational Measurement: Issues and Practice*, 16 (2), 16–18

Meisels, S J, Atkins-Burnett, S, Xue, Y, Nicholson, J, Bickel, D D & Son, S (2003) Creating a system of accountability: The impact of instructional assessment on elementary children's achievement test scores, *Education Policy Analysis Archives*, 11 (9) http://epaa.asu.edu/ojs/article/viewFile/237/363

Mesic, V (2012) Identifying Country-Specific Cultures of Physics Education: A differential item functioning approach, *International Journal of Science Education*, 34 (16), 2483–2500

Messick, S (1965) Personality measurement and the ethics of assessment, *American Psychologist*, 20 (2), 136–142

Messick, S (1980) Test validity and the ethics of assessment, *American Psychologist*, 35 (11), 1012–1027

Messick, S (1981) Evidence and ethics in the evaluation of tests, *Educational Researcher*, 10 (9), 9–20

Messick, S (1988) The once and future issues of validity: assessing the meaning and consequences of measurement. In: H Wainer & H I Braun (eds.) *Test Validity.* Hillsdale, NJ, Lawrence Erlbaum Associates, Inc. 33–45

Messick, S (1989a) Validity. In: R Linn (ed.) *Educational Measurement* (Third edition). Washington, DC, American Council on Education, 13–100

Messick, S (1989b) Meaning and values in test validation: The science and ethics of assessment, *Educational Researcher*, 18 (2), 5–11

Messick, S & Anderson, S (1974) Educational testing, individual development, and social responsibility. In: R W Tyler & R M Wolf (eds.) *Crucial Issues in Testing*. Berkeley, CA, McCutchan Publishing Corporation, 21–34

Michell, J (1997) Quantitative Science and the Definition of Measurement in Psychology, *British Journal of Psychology*, 88 (3), 355–383

Michell, J (1999) *Measurement in Psychology: A Critical History of a Methodological Concept*. Cambridge, Cambridge University Press.

Michell, J (2000) Normal science, pathological science and psychometrics, *Theory and Psychology*, 10 (5), 639–667

Michell, J (2009) Invalidity in validity. In: R W Lissitz (ed.) *The Concept of Validity: Revisions, New Directions, and Applications*. Charlotte, NC, Information Age Publishing, 111–133

Mislevy, R J (1997) *Postmodern test theory. Transitions in Work and Learning: Implications for Assessment*. National Academy of Sciences. Washington, DC, National Academies Press

Mislevy, R J (2003a) *Argument Substance and Argument Structure in Educational Assessment*. CSE Report 605. LA, California: National Center for Research on Evaluation, Standards, and Student Testing, University of California.

Mislevy, R J (2003b) Substance and structure in assessment arguments, *Law, Probability, and Risk*, 2 (4), 237–258

Mislevy, R J (2009) Validity from the perspective of model-based reasoning. In: R W Lissitz (ed.) *The Concept of Validity: Revisions, New Directions, and Applications*. Charlotte, NC, Information Age Publishing, 83–108

Mislevy, R J, Steinberg, L S & Almond, R G (2003) On the structure of educational assessments, *Measurement: Interdisciplinary Research and Perspectives*, 1 (1), 3–62

Mittler, P J (1973) Purposes and principles of assessment. In: P J Mittler (ed.) *Assessment for learning in the mentally handicapped*. London, Churchill Livingstone

Moe, O (2002) 'En erfaring rikere' – Ekstern deltakelse i lokalt. vurderingsarbeid. In: P Haug & L Monsen (eds.) *Skolebasert vurdering – erfaringer og utfordringer*. Oslo, Abstakt forlag, 57–77

Monroe, W S & Souders, L B (1923) *The Present State of Written Examinations and Suggestions for their Improvement*. Urbana-Champaign, IL, University of Illinois

Monseur, C & Berenzer, A (2007) The computation of equating errors in international surveys in education*, Journal of Applied Measurement*, 8 (3), 323–335

Monseur, C & Adams R (2009) Plausible values: how to deal with their limitations*, Journal of Applied Measurement*, 10 (3), 320–334

Moran, M, Rein, M & Goodin, R E (2008) *The Oxford Handbook of Public Policy*. Oxford, Oxford University Press

Moss, P A (1992) Shifting conceptions of validity in educational measurement: implications for performance assessment, *Review of Educational Research*, 62 (3), 229–258

Moss, P A (1994) Can there be validity without reliability?, *Educational Researcher*, 23 (2), 5–12

Moss, P A (1995) Themes and variations in validity theory, *Educational Measurement: Issues and Practice*, 14 (2), 5–13

Moss, P A (1996) Enlarging the dialogue in educational measurement: Voices from interpretive research traditions, *Educational Researcher*, 25 (1), 20–28

Moss, P A (1998) The role of consequences in validity theory, *Educational Measurement: Issues and Practice*, 17 (2), 6–12

Moss, P A (2003) Reconceptualizing validity for classroom assessment, *Educational Measurement: Issues and Practice*, 22 (4), 13–25

Moss, P A (2013) Validity in action: Lessons from studies of data use, *Journal of Educational Measurement*, 50 (1), 91–98

Moss, P A, Girard, B J & Haniford, L C (2006) Validity in educational assessment, *Review of Research in Education*, 30, 109–162

Moss, P A & Koziol, S M (1991) Investigating the validity of a locally developed critical thinking test, *Educational Measurement: Issues and Practice*, 10 (3), 17–22

Moss, P A, Pullin, D C, Gee, J P & Haertl, E H (2005) The idea of testing: Psychometric and sociocultural perspectives, *Measurement: Interdisciplinary Research and Perspectives*, 3 (2), 63–83

Moss, P A, Pullin, D C, Gee, J P , Haertl, E H & Young, L J (2005) *Assessment, Equity and Opportunity to Learn*. Cambridge, Cambridge University Press

Mullis, I V S, Martin, M O, Foy, P & Arora, A (2012) *TIMSS 2011 international results in mathematics.* Chestnut Hill, MA, TIMSS and PIRLS International Study Center, Boston College. http://timssandpirls.bc.edu/timss2011/international-results-mathematics.html

Mullis I V S, & Martin (2013) *PIRLS 2016 Assessment Framework*. TIMSS and PIRLS International Study Center, Boston College

Nardi, E (2008) Cultural biases: a non-Anglophone perspective, *Assessment in Education: Principles, Policy & Practice*, 15 (3), 259–266

Natriello, G (1987) The impact of evaluation processes on students, *Educational Psychologist*, 22(2), 155–175

Newton, P E (2005) The public understanding of measurement inaccuracy, *British Educational Research Journal*, 31 (4), 419–442

Newton, P E (2007a) *Evaluating assessment systems*. Paper 1. Submission to the Education and Skills Select Committee Inquiry into Testing and Assessment. London, Qualifications and Curriculum Authority
https://orderline.education.gov.uk/gempdf/1445900599/Evaluating_Assessment_Systems1.pdf

Newton, P E (2007b) Clarifying the purposes of educational assessment, *Assessment in Education: Principles, Policy & Practice*, 14 (2), 149–170

Newton, P E (2009) The reliability of results from national curriculum testing in England, *Educational Research*, 51 (2), 181–212

Newton, P E (2012a) Clarifying the consensus definition of validity, *Measurement: Interdisciplinary Research and Perspectives*, 10 (1–2), 1–29

Newton, P E (2012b) Validity, purpose and the recycling of results from educational assessments. In: J Gardner (ed.) *Assessment and Learning* (Second edition). London, SAGE, 264–276

Newton, P E & Shaw, S D (forthcoming, 2014) *Validity in educational and psychological assessment*. London, SAGE

Nichols, P D, Meyers, J L & Burling, K S (2009) A framework for evaluating and planning assessments intended to improve student achievement, *Educational Measurement: Issues and Practice*, 28 (3), 14–23

Norris, S P (1995) Measurement by tests and consequences of test use. *Philosophy of Education*. Urbana, IL, Philosophy of Education Society, 303–306

Novoa A & Yariv-Marshal T (2003) Comparative research in education: a mode of governance or a historical journey? *Comparative Education*, 39(4), 423 – 438

Nusche, D, Earl, L, Maxwell, W & Shewbridge, C (2011) *OECD Reviews of Evaluation and Assessment in Education*. Organisation for Economic Cooperation and Development (www.oecd.org/norway/48632032.pdf)

Nuttall, D L (1987) The validity of assessments, *European Journal of Psychology of Education*, 2 (2), 109–118

Nyquist, J B (2003) *The benefits of reconstructing feedback as a larger system of formative assessment: A meta-analysis*. Unpublished master's thesis. Nashville, TN, Venderbuilt University

Østerud, (2006) PISA- og TIMSS- undersøkelsene. Hvor viktige er de for norsk skole, og hvilke lærdommer kan vi høste? In: B Brock-Utne & L Bøyesen (eds.) Aa greie seg i utdanningssystemet i nord og sør. Innføring i flerkulturell og komparativ pedagogikk, utdanning og utvikling. Fagbokforlaget, Bergen, 204–220

Oancea, A (2005) Criticisms of educational research: key topics and levels of analysis, *British Educational Research Journal*, 31 (2), 157–183

OECD (2010) *Viewing the United Kingdom School System through the Prism of PISA* (www.oecd.org/pisa/46624007.pdf)

OECD (2012) *PISA 2009 Technical Report*(www.oecd.org/pisa/pisaproducts/50036771.pdf)

OECD (2013) *PISA 2012 Results in Focus: What 15-Year-olds know and what they can do with what they know* (www.oecd.org/pisa/keyfindings/pisa-2012-results-overview.pdf)

Ofqual (2012) *GCSE English 2012.* Ofqual/12/5225. Coventry, Office of Qualifications and Examinations Regulation

Ofqual (2013a) *Review of Controlled Assessment in GCSEs*. Coventry, Office for Qualifications and Examinations Regulation

Ofqual (2013b) *Reforms to GCSEs in England from 2015*. Coventry, Office for Qualifications and Examinations Regulation

Ofsted (2008) *Assessment for learning: the impact of National Strategy support*. London, Office for Standards in Education

Oliveri, M E & Ercikan, K (2011) Do Different Approaches to Examining Construct Comparability in Multilanguage Assessments Lead to Similar Conclusions?, *Applied Measurement in Education*, 24 (4), 349-366

O'Neill, O (2005) *Assessment, public accountability and trust* (www.cambridgeassessment.org.uk)

Orton, R E (1998) Samuel Messick's consequential validity. *Philosophy of Education*. Urbana, IL, Philosophy of Education Society, 538–545

Ozga, J (2012) Introduction. Assessing PISA, *European Educational Research Journal*, 11 (2), 166–171

Pellegrino, J W, Chudowsky, N & Glaser, R (eds.) (2001) *Knowing what students know: The science and design of educational assessment*. Washington, DC, National Academies Press

Pellegrino, J W & Hilton, M L (eds.) (2012) *Committee on Defining Deeper Learning and 21st Century Skills*. Center for Education, Division on Behavioral and Social Sciences and Education, National Research Council

Perie, M, Marion, S & Gong, B (2007) *A framework for considering interim assessments.* National Center for the Improvement of Educational Assessment. Dover, NH, NCIEA (www.nceia.org)

Perry, N, Hutchinson, L & Thauberger, C (2007) Mentoring Student Teachers to Design and Implement Literacy Tasks That Support Self-Regulated Reading and Writing, *Reading and Writing Quarterly*, 23 (1), 27–50

Perrenoud, P (1998) From formative evaluation to a controlled regulation of learning processes: Towards a conceptual field, *Assessment in Education: Principles, Policy & Practice*, 5 (1), 85–102

Phelps, R P (2012) The Effect of Testing on Student Achievement, 1910–2010, *International Journal of Testing*, 12 (1), 21–43

Phillips, D C (2014) Research in the hard sciences, and in the very hard 'softer' domains, *Educational Researcher*, 43 (1), 9–12

Piaget, J (1967) *The child's conception of the world*. Totowa, NJ, Littlefield, Adams

Polesel, J, Dulfer, N & Turnbull, M (2012) *The Experience of Education: The impacts of high stakes testing on school students and their families.* Literature Review. Sydney, NSW, The Whitlam Institute, University of Western Sydney

Popham, W J (1987) The merits of measurement-driven instruction, *Phi Delta Kappan*, 68 (9), 679–82

Popham, W J (1997) Consequential validity: Right concern – wrong concept, *Educational Measurement: Issues and Practice*, 16 (2), 9–13

Popham, W J (2007). Instructional sensitivity of tests: Accountability's dire drawback, *Phi Delta Kappan*, 89 (2), 146-150

Popham, W J (2008a) *Transformative Assessment.* Association for Supervision and Curriculum Development (ASCD), Alexandria, VA

Popham, W J (2008b) *Classroom assessment: What teachers need to know* (5th ed). Boston, Prentice Hall

Popper, K R (1959) *The logic of scientific discovery*. London, Hutchinson Education

Porter, T M (1986) *The rise of statistical thinking. 1820–1900.* Princeton, NJ, Princeton University Press

Powell, W W & Snellman, K (2004) The Knowledge Economy, *Annual Review of Sociology*, 30, 199–220

Power, M (1999) *The Audit Society: Rituals of Verification*. Oxford, Oxford University Press

Power, M (2000) The Audit Society – Second Thoughts, *International Journal of Auditing*, 4, 111–119

Prais, S J (2004) Cautions on OECD's recent educational survey (PISA), *Oxford Review of Education*, 29 (2), 139–163

Prais, S J (2007) Two recent (2003) international surveys of schooling attainments in mathematics: England's problems, *Oxford Review of Education*, 33 (1), 33–46

Pring, R, Hayward, G, Hodgson, A, Johnson, J, Keep, E, Oancea, A, Rees, G, Spours, K & Wilde, S (2009) *Education for All. The Future of Education and Training for 14–19 Year-Olds*. London, Routledge

Pryor, J & Crossouard, B (2008) A socio-cultural theorisation of formative assessment, *Oxford Review of Education*, 34 (1), 1–20

Puk, T (1999) Formula for success according to TIMSS or the subliminal decay of jurisdictional educultural identity?, *Alberta Journal of Educational Research*, 45 (3), 225–238

QCA (2005) *A Review of GCE and GCSE Coursework Arrangements*. London, Qualifications and Curriculum Authority

QCA (2006) *A Review of GCSE coursework*. London, Qualifications and Curriculum Authority

QCA (2007) *Controlled Assessments*. London, Qualifications and Curriculum Authority

Ramaprasad, A (1983) On the definition of feedback, *Behavioral Science*, 28, 4–13

Rambøll (2013) *Evaluering av årlige tilstandsrapporter*. Delrapport 2. Rapport til Utdanningsdirektoratet

Rasch, G (1960) *Probabilistic Models for Some Intelligence and Achievement Tests*. Copenhagen, Danish Institute for Educational Research. Expanded edition (1983), Chicago, MESA Press

Rautalin, E & Alasuutari, P (2008) The uses of the national PISA results by Finnish officals in central government, *Journal of Education Policy*, 24 (5), 539–556

Reckase, M D (2001) Innovative methods for helping standard setting participants to perform their task: The role of feedback regarding consistency, accuracy, and impact. In: G C Cizek (ed.) *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ, Lawrence Erlbaum Associates Inc, 159–173

Resnick, L B & Resnick, D P (1992) Assessing the thinking curriculum: New tools for educational reform. In: B R Gifford & M C O'Connor (eds.) *Changing Assessments: Alternative Views of Aptitude, Achievement, and Instruction*. Boston, MA, Kluwer, 37–75

Rinne, R (2008) The growing supranational impacts of the OECD and the EU on national education policies and the case of Finland, *Policy Futures in Education*, 6, 665–680

Rodriguez, M C (2004) The role of classroom assessment in student performance on TIMSS, *Applied Measurement in Education*, 17 (1), 1–24

Roos, B & Hamilton, D (2005) Formative assessment: a cybernetic viewpoint, *Assessment in Education: Principles, Policy & Practice*, 12 (1), 7–20

Ruch, G M (1924) *The Improvement of the Written Examination*. Chicago, Scott, Foresman and Company

Rust, J & Golombok, S (1989) *Modern Psychometrics. The science of psychological assessment* (Second edition). Reprinted 2000. London, Routledge

Sadler, R D (2007) Perils in the meticulous specification of goals and assessment criteria, *Assessment in Education: Principles, Policy & Practice*, 14 (3), 387–392

Sadler, D R (1989) Formative Assessment and the Design of Instructional Systems, *Instructional Science,* 18, 119–144

Sadler, D R (2010) Beyond feedback: developing student capability in complex appraisal, *Assessment & Evaluation in Higher Education,* 35 (5), 535–550

Sadler, D R (2005) Interpretations of criteria-based assessment and grading in higher education, *Assessment & Evaluation Higher Education*, 30 (2), 175–194

Sandilands, D, Oliveri, M E, Zumbo, B D & Ercikan, K (2013) Investigating Sources of Differential Item Functioning in International Large-Scale Assessments Using a Confirmatory Approach, *International Journal of Testing*, 13 (2), 152–174

Sato, M et al. (2005) Two teachers making assessment for learning their own, *Curriculum Journal,* 16 (2), 177-191

Schleicher, A (2006) *The economics of knowledge: Why education is key for Europe's success.* Policy Brief. The Lisbon Council

Schleicher, A (2009) International Benchmarking as a Lever for Policy Reform. In: A Hargreaves & M Fullan (eds.) *Change Wars*. Bloomington, IN, Solution Tree

Schnepf, S V (2007) Immigrants' educational disadvantage: an examination across ten countries and three surveys, *Journal of Population Economics,* 20, 527–545

Shulman, L S (1986) Those who understand: Knowledge growth in teaching, *Educational Researcher*, 15 (2), 4–14

Schwippert, K & Lenkeit, J (eds.) (2012) *Progress in Reading Literacy in National and International Context.* Studies in International Comparative and Multicultural Education, Vol 13, The Impact of PIRLS 2006 in 12 countries. Munster, Waxman

Scriven, M (1967) The methodology of evaluation. In: R Tyler, R Gagne & M Scriven (eds.) *Perspectives on Curriculum Evaluation*. AERA Monograph Series – Curriculum Evaluation. Chicago, Rand McNally and Co

Searle, J R (1990) Is the brain a digital computer? *Proceedings and Addresses of the American Philosophical Association*, 64 (3), 21–37 (www.jstor.org/stable/3130074)

Sfard, A (1998) On two metaphors for learning and the dangers of choosing just one, *Educational Researcher*, 27 (2) 4–13

Shaw, S & Crisp, V (2012) An approach to validation: Developing and applying an approach for the validation of general qualifications, *Research Matters*, Special Issue, 3, 1–44

Shepard, L (1992) Will national tests improve student learning? CSE Technical Report 342. CRESST, University of Colorado, Boulder

Shepard, L A (1997a) *Measuring achievement: What does it mean to test for robust understanding?* William H Angoff Memorial Lecture Series. Princeton, NJ, Educational Testing Service

Shepard, L A (1997b) The centrality of test use and consequences for test validity, *Educational Measurement: Issues and Practice*, 16 (2), 5–8

Shepard, L A (2000) The role of assessment in a learning culture, *Educational Researcher*, 29 (7), 4–14

Shute, V J (2008) Focus on formative feedback, *Review of Educational Research*, 78 (1), 153–189

Sijtsma, K (2006) Psychometrics in Psychological Research: Role Model or Partner in Science? *Psychometrika*, 71 (3), 451–455

Simon, H A (1957) *Models of Man.* New York, John Wiley and Sons

Skarheim, P (2010) *Assessment Policies and Research in Norway*. Keynote by Petter Skarheim, Norwegian Directorate for Education and Training. Association for Educational Assessment in Europe, 11th Annual Conference. Oslo, Norway

Skedsmo, G (2011) Formulation and realisation of evaluation policy: inconsistencies and problematic issues, *Journal of Educational Assessment, Evaluation and Accountability*, 23 (1), 5–20

Skinner, B F (1938) *The behaviour of organisms*. New York, Appleton-Century-Crofts

Skinner, B F (1950) Are theories of learning necessary? *The Psychological Review*, 57 (4)

Skinner, B F (1989) Teaching machines, *Science*, New Series, 128 (3330), 969–977

Smith, E & Gorard, S (2005) 'They don't give us our marks': the role of formative feedback in student progress, *Assessment in Education: Principles, Policy & Practice*, 12 (1), 21–38

Smith, K (2007) Vurdering som et motivasjonsfremmende redskap for læring, *Norsk Pedagogisk Tidsskrift*, 2, 100–105

Solano-Flores G, Backhoff E & Contreras-Nino, L A (2009) Theory of test translation error, *International Journal of Testing*, 9 (2), 78–91

Spearman, C (1904) "General intelligence". Objectively determined and measured, *Journal of Psychology*, 1 (15), 201–293

Stanat, P & Lüdtke, O (2008) Multilevel issues in international large-scale assessment studies on student performance. In: F J R van de Vijver, D A van Hemert & Y H Poortinga, Y H (eds.) *Individuals and cultures in multilevel analysis*. Hillsdale, NJ, Erlbaum, 315–344

Stanat, P & Lüdtke, O (2013) International Large-Scale Assessment Studies of Student Achievement. In: J. Hattie & E M Anderman (eds.) *International guide to Student Achievement*. Educational Psychology Handbook Series. Alexander, P (series ed.). New York, Routledge, Taylor and Francis

Stevens, S S (1946) On the Theory of Scales of Measurement, *Science*, 103 (2684), 677–680

Steiner-Khamsi, G (Ed) (2004) *The Global Politics of Educational Borrowing and Lending*, New York, Teachers College Press

Stiggins, R J (1999) Evaluating classroom asssessment training in teacher education programs. *Educational Measurement: Issues and Practice*, 18 (1), 23-27.

Stiggins, R J (2005) *Student-involved assessment for learning.* Upper Saddle River, NJ, Prentice Hall

Stiggins, R & Arter, J (2002) *Assessment for Learning: International Perspectives*. The Proceedings of an International Conference. Chester, United Kingdom, 18 September 2001. Corporate source: Assessment Training Inst., Inc., Portland

Stobart, G (2001) The validity of national curriculum assessment, *British Journal of Educational Studies*, 49 (1), 26–39

Stobart, G (2002) Using Assessment To Improve Learning: Intentions, Feedback And Motivation. In: C Richardson (ed.) *Whither Assessment?* London, QCA, 113-122

Stobart, G (2006) The validity of formative assessment. In: J Gardner (ed.) *Assessment and Learning*. London, SAGE, 133–146

Stobart, G (2008) *Testing Times: The uses and abuses of assessment.* Oxon, Routledge, Taylor and Francis

Stobart, G (2009) Determining validity in national curriculum assessments, *Educational Research*, 51 (2), 161–179

Stobart, G (2014) *The Expert Learner. Challeging the Myth of Ability*. L Stoll & L Earl (series eds.) Expanding Educational Horizons. McGraw-Hill, Open University Press

Stockford, I, Meadows, M, Wheadon, C, Pinot de Moira, A, Eason, S, Taylor, M & Faulkner, K (2012) *AQA GCSE English and GCSE English Language January 2011 to June 2012: Report to Ofqual*. Manchester, Assessment and Qualifications Alliance

Strietholt, R et al. (2013) A correction model for differences in the sample compositions: the degree of comparability as a function of age and schooling, *Large-scale Assessments in Education* 1 (1)

Suchan, B, Wallner-Paschon, C & Riess, C (2012) The Impact of PIRLS in Austria. In: K Schwippert & J Lenkeit (eds.) *Progress in Reading Literacy in National and International Context*. Studies in International Comparative and Multicultural Education, Vol 13, The Impact of PIRLS 2006 in 12 countries. Munster Waxman

Suppe, F (1984) Beyond Skinner and Kuhn, *New Ideas in Psychology*, 2 (2), 89–104

Takayama, K (2008) The politics of international league tables: PISA in Japan's achievement crisis debate, *Comparative Education*, 44 (4), 387–407

Takayama, K (2010) Politics of externalization in reflexive times: reinventing Japanese education reform discourses through 'Finnish PISA success', *Comparative Education Review*, 54 (1), 51–75

Tapan, S (2001) *A study of present learning assessment practices to ensure quality of primary education*. Basic education studies, 2000-2001. Dhaka, UNESCO Office Dhaka

Taras, M (2009) Summative assessment: the missing link for formative assessment, *Journal of Further and Higher Education*, 33 (1), 57–69

Telhaug, A O et al. (2006) The Nordic model in education: Education as part of the political system in the last 50 years, *Scandinavian Journal of Educational Research,* 50 (3), 245–283

Telhaug, A O (2007) Kunnskapsløftet i et utdanningshistorisk perspektiv. In: I H Hølleland (ed.) *På vei mot Kunnskapsløftet. Begrunnelser, løsninger og utfordringer.* Oslo, Cappelen Akademisk Forlag, 47–65

Thompson, M & Wiliam, D (2008) Tight but Loose: A Conceptual Framework for Scaling Up School Reforms. In: C Wylie (ed.) *Tight but Loose: Scaling Up Teacher Professional Development in Diverse Contexts*. Princeton, NJ, Educational Testing Service

Tierney, R & Charland, J (2007) *Stocks and Prospects: Research on Formative Assessment in Secondary Classrooms*. Online Submission, Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 11 April

Torrance (2007) Assessment as learning? How the use of explicit learning objectives, assessment criteria and feedback in post secondary education and training can come to dominate learning, *Assessment in Education: Principles, Policy & Practice*, 14 (3) 281–294.

Torrance, H (2000) Postmodernism and Educational Assessment. In: A Filer (ed.) *Assessment. Social Practice and Social Product*. London, Routledge, 173–188

Torrance, H & Pryor, J (1998) Investigating Formative Assessment: Teaching; Learning and Assessment in the Classroom, *British Journal of Educational Studies,* 47 (3), 284–285

Torrance, H & Pryor, J (1998) *Investigating Formative Assessment. Teaching, Learning and Assessment in the Classroom*. Buckingham, Open University Press

Torrance, H. & Pryor, J. (2001) Developing formative assessment in the classroom: using action research to explore and modify theory, *British Educational Research Journal*, 27 (5), 615–631

Thune, T, Kyvik, S, Sörlin, S, Olsen, T B, Vabø, A & Tømte, C (2012) *PhD education in a knowledge society. An evaluation of PhD education in Norway*. Oslo, NIFU

Välijärvi, J, Linnakylä, P, Kupari, P, Reinikainen, P & Arffman, I (2002) *The Finnish success in PISA – And some reasons behind it*. Jyväskylä: Institute for Educational Research, University of Jyväskylä

Vygotsky, L S (1978) *Mind in Society*. Cambridge, MA, Harvard University Press

Wagemaker, H (2008) Choices and tradeoffs: reply to McGaw, *Assessment in Education: Principles, Policy & Practice*, 15 (3), 267–278. Special Issue: International Comparative Studies in Achievement

Watson, A (2006) Some Difficulties in Informal Assessment in Mathematics, *Assessment in Education: Principles, Policy & Practice*, 13 (3), 289–303

Welle-Strand, A & Tjeldvold, A (2002) The Norwegian Unified School – A Paradise Lost?, *Journal of Education Policy*, 17 (6), 673–686

Webb, M & Jones, J (2009) Exploring tensions in developing assessment for learning, *Assessment in Education: Principles, Policy & Practice*, 16 (2), 165–184

Westbury, I (1992) Comparing American and Japanese Achievement: is the United States really a low achiever?, *Educational Researcher*, 21 (5), 18–24

Whetton, C, Twist, L & Sainsbury M (2007) Measuring standards in primary English: the validity of PIRLS - a response to Mary Hilton, *British Educational Research Journal*, 33 (6), 977–986

Wieman, C E (2014) The similarities between research in education and in the hard sciences, *Educational Researcher*, 43 (1), 12–14

Wiley, D E (1991) Test validity and invalidity reconsidered. In: R E Snow & D E Wiley (eds.) *Improving Inquiry in Social Science: A Volume in Honor of Lee J. Cronbach*. Hillsdale, NJ:, Lawrence Erlbaum Associates, Inc, 75–107

Wiliam, D, Lee, C, Harrison, C, & Black, P (2004) Teachers Developing Assessment for Learning: Impact on Student Achievement, *Assessment in Education: Principles, Policy & Practice,*. 11 (1), 49–65

Wiliam, D (2007) Keeping learning on track: Classroom assessment and the regulation of learning. In: F K Lester Jr (ed.) *Second handbook of mathematics teaching and learning*. Greenwich, CT, Information Age Publishing, 1053–1098

Wiliam, D (2008) International comparisons and sensitivity to instruction, *Assessment in Education: Principles, Policy a&nd Practice*, 15 (3), 253 – 257

Wiliam, D (2011) What is assessment for learning?, *Studies in Educational Evaluation*, 37 (1), 2–14

Wiliam, D & Black, P (1996) Meanings and consequences: a basis for distinguishing formative and summative functions of assessment, *British Educational Research Journal*, 22 (5), 537–548

Willis, J (2008) Assessment for learning: a socio-cultural approach. In: 'AARE 2008 International education research conference: Brisbane: papers collection' (Conference of the Australian Association for Research in Education, 30 November – 4 December 2008) compiled by P L Jeffrey. Melbourne, Australian Association for Research in Education

Wilson, N (1998) Educational standards and the problem of error, *Education Policy Analysis Archives*, 6 (10)

Wilson, M (2005) *Constructing measures: an item response modelling approach*. Mahwah, NJ, Erlbaum

Wilson M & Scalise K (2006) Assessment to improve learningin Higher Education: The BEAR Assessment System*, Higher Education*, 52 (4), 635–663

Wilson, M (2009) Measuring Progressions: Assessment Structures Underlying a Learning Progression, *Journal of Research in Science Teaching*, 46 (6), 716–730

Wolf, L F, Smith, J K & Birnbaum, M E (1995) Consequence of performance, test, motivation, and mentally taxing items, *Applied Measurement in Education*, 8 (4), 341–351

Wolpert, L (1992) *The unnatural nature of science. Why science does not make (common) sense.* London, Faber and Faber, 56–85

Wu, M (2005) The role of plausible values in large-scale surveys, *Studies in Educational Evaluation*, 31, 114–128

Wu, M (2006) Using multiple-variable matching to identify cultural sources of differential item functioning, *International Journal of Testing*, 6 (3), 287–300

Zumbo, B D (2009) Validity as contextualized and pragmatic explanation, and its implications for validation practice. In: R W Lissitz (ed.) *The Concept of Validity: Revisions, New Directions, and Applications*. Charlotte, NC, Information Age Publishing, 65–82

Zumbo, B D & Forer, B (2011) Testing and measurement from a multilevel view: psychometrics and validation. In: J A Bovaird, K F Geisinger & C W Buckendahl (eds.) *High-Stakes Testing in Education: Science and Practice in K-12 Settings*. Washington, DC, American Psychological Association, 177–190

About the authors

Professor Jo-Anne Baird is the Pearson Professor of Educational Assessment and Director of OUCEA. She is Lead Editor of the international research journal *Assessment in Education: Principles, Policy and Practice*. Jo-Anne previously held the position of Head of Research at the Assessment and Qualifications Alliance, where she managed the research programme and was responsible for the standard-setting systems for public examinations. Her research interests include examination standards, e-assessment and human judgement in assessment. She is President of the Association for Educational Assessment – Europe. Together with Paul Black, Jo-Anne is a Guest Editor of a special issue of Research Papers in Education, arising from work conducted for the Reliability Programme of the Office of Qualifications and Examinations Regulation (Ofqual). Jo-Anne directed the 14–19 Centre Research Study, which investigated the impact of assessment and educational reforms upon 52 schools and colleges in England.

Selection of relevant publications

- Baird, J & Black, P (2013) Test theories, educational priorities and reliability of public examinations in England. *Research Papers in Education*, 28 (1), 5–21.
- Rose, J & Baird, J (2013) Aspirations and an austerity state: young people's hopes and goals for the future. *London Review of Education*, 11 (2), 157–173.
- Daly, A, Baird, J, Chamberlain, S & Meadows, M (2012) Assessment reform: students' and teachers' responses to the introduction of stretch and challenge at A-level. *The Curriculum Journal*, 23 (2), 139–155.
- Baird, J, Isaacs, T, Johnson, S, Stobart, G, Yu, G, Sprague, T & Daugherty, R (2011) *Policy Effects of PISA*. Report commissioned by Pearson UK.

Dr Therese Hopfenbeck is Lecturer in Educational Assessment at OUCEA and conducts research on international tests, especially the relationship between student motivation and results on the tests. She is also interested in classroom assessment and self-regulation in student learning. Therese has recently been commissioned by the Norwegian government and OECD to investigate assessment policy implementation in Norway.

Selection of relevant publications

- Hopfenbeck, T, Tolo, A, Florez, T & El Masri, Y (2013) *Balancing Trust and Accountability? The Assessment for Learning Programme in Norway*. A Governing Complex Education Systems Case Study. Paris: Organisation for Economic Co-operation and Development.

- Hopfenbeck, T N, Throndsen, I, Lie, S & Dale, E L (2012) Assessment with distinctly defined criteria: A research study of a national project. *Policy Futures in Education*, 10 (4), 421–433.
- Hopfenbeck, T N, Kjærnsli, M & Olsen, R V (eds.) (2012) *Kvalitet i norsk skole. Internasjonale og nasjonale undersøkelser av læringsutbytte og undervisning*. (Quality in the Norwegian school. International and national tests of learning outcomes and teaching). Oslo: Universitetsforlaget.
- Hopfenbeck, T N (2013) What did you learn in school today? In Hattie, J, Wille, T S, Hermansen, M, Hopfenbeck, T N, Madsen, C, Kirkegaard, P, Bjerresgaard, H, Weinstein, CE, Bråten, I & Andreassen, R (eds.) *Feedback og vurdering for læring*, in Danish.

Professor Paul Newton

Paul Newton is interested in all aspects of educational assessment, although the focus of his work is often upon large-scale summative assessment systems. He explores the conceptual foundations of educational assessment, and is currently developing a programme of work focused on validity and evaluation frameworks within educational and psychological measurement. Other interests include the comparability of examination standards, the uses to which educational assessment results are put, the history of educational and psychological measurement, and the public understanding of educational assessment. Paul has extensive experience of research and policy analysis within national assessment agencies, having worked for a range of UK examining boards and regulatory authorities. Paul is an Executive Editor for *Assessment in Education: Principles, Policy and Practice.*

Selection of relevant publications

- Newton, P E & Shaw, S D (2014, in press). *Validity in educational and psychological assessment*. London: SAGE.
- Newton, P E (2012) Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives*, 10 (1–2), 1–29.
- Newton, P E (2010) Thinking about linking. *Measurement: Interdisciplinary Research and Perspectives*, 8(1), 38–56.
- Newton, P E (2007) Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy and Practice*, 14 (2), 149–170.

Professor Gordon Stobart

After teaching for eight years in secondary schools in Africa and Inner London, Gordon became an Educational Psychologist. This led to a Fulbright scholarship in the USA, where he gained a PhD for research into the integration of special needs students in mainstream classrooms.

On his return to the UK, Gordon worked as an assessment researcher for 15 years, firstly with an exam board and then with government agencies. These posts led to wide experience with assessment policy and the development of national qualifications and assessments. Gordon moved to the Institute of Education, University of London, where he focused on the formative role of assessment. As a member of the Assessment Reform Group, he worked for over a decade on developing AfL, an approach which now has international recognition. He has recently been working on the acclaimed national implementation of *Vurdering for Læring.*[26]

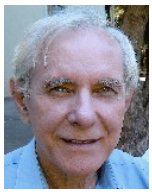
Selection of relevant publications

- Stobart, G (2014) *The Expert Learner. Challenging the Myth of Ability*. Open University Press.
- Stobart, G & Eggen, T (2012) High-stakes testing – value, fairness and consequences. *Assessment in Education: Principles, Practice and Policy*, 19 (1), 1–6.
- Stobart, G (2012) Validity in formative assessment. In Gardner, J (ed.) *Assessment and Learning* (Second edition). London: SAGE.
- Stobart, G (2009) Determining validity in national curriculum assessments. *Educational Research*, 51 (2), 161–173.
- Stobart, G (2008) *Testing Times*. Abingdon, UK: Routledge.

Anna Therese Steen-Utheim is a PhD student in educational assessment and learning at the University of Oslo and a senior adviser at the LearningLab, the Norwegian Business School. The working title of her PhD is 'Assessment and learning; feedback on written and oral work and its impact on students learning'.

[26] www.udir.no/Vurdering/Vurdering-for-laring.

The advisory group



Professor David Andrich's current research in applying Rasch models for measurement has two strands. The first involves articulating a research and assessment paradigm that is different from the traditional in which statistical models are applied. In the traditional paradigm, the case for choosing any model to summarise data is that it fits the data at hand; in contrast, in applying the paradigm of Rasch models, the case for these models is that if the data fit the model, then, within a frame of reference, they provide invariance of comparisons of persons with respect to items, and vice versa. Then any misfit between the data and the chosen Rasch model is seen as an anomaly that needs to be explained qualitatively by reference to the theory behind the construction of the instrument, as well as the operational aspects of its application. David argues that this approach improves the quality of social measurement, including in education, psychology, sociology, economics and health outcomes. The second area of research is further articulating the implications of the Rasch models and development of complementary software, to better understand a range of anomalies – for example, how to identify guessing in multiple-choice items, how to identify and handle response dependence between items, and multidimensionality.

Selection of relevant publications

- Andrich, D (2010) Educational Measurement: Rasch Models. In Baker, E, Peterson, P & McGaw, B (eds.) *International Encyclopedia of Education* (Third edition). Elsevier.
- Andrich, D (2004) Controversy and the Rasch Model: A Characteristic of Incompatible Paradigms? *Medical Care*, 42 (1), 17–116, January Supplement: Applications of Rasch Analysis in Health Care. Lippincott Williams and Wilkins.
- Andrich, D (2003) On the distribution of measurements in units that are not arbitrary. *Social Science Information*, 42, 4, 557–589.
- Andrich, D (2002) A framework relating Outcomes Based Education and the Taxonomy of Educational Objectives. *Studies in Educational Evaluation*, 28, 35–59.

Professor Mary James retired from the University of Cambridge in December 2013. She is a non-executive director of Bell Educational Services, which runs English language courses, and Chair of the Scientific Advisory Board for the NordForsk programme, 'Education for Tomorrow'. She was President of the British Educational Research Association until 2013. Her research interests encompass curricula, pedagogy and assessment in schools, along with implications for teachers' professional development, school leadership and policy frameworks. She was a Lecturer, Senior Lecturer, and then Reader in the University of Cambridge School of Education from 1989 until becoming a Chair of

Education at the Institute of Education, University of London (2005–2008). Previously she was a Research Fellow at the Open University. She began her career by teaching religious education, English and social studies in three secondary schools. Mary's selected works have just been published by Routledge. She is the first female whose work has been selected for publication in the World Library of Educationalists series.

Selection of relevant publications

- James, M E (2013) *Educational Assessment, Evaluation and Research: The selected works of Mary E. James.* In the World Library of Educationalists. Abingdon, UK: Routledge.
- James, M (2012) Assessment in harmony with our understanding of learning: problems and possibilities. In Gardner, J (ed.) *Assessment and Learning* (Second edition). London: SAGE, 187–205.
- Daugherty, R, Black, P, Ecclestone, K, James, M & Newton, P (2011) Assessment of Significant Learning Outcomes. In Berry, R & Adamson, R (eds.) *Assessment Reform in Education*. New York: Springer, 165–183.
- James, M (2011) Assessment for Learning: research and policy in the (dis)United Kingdom. In Berry, R & Adamson, R (eds.) *Assessment Reform in Education*. New York: Springer, 15–32.

Professor Dylan Wiliam's principal research focus is the professional development of teachers through a focus on the use of evidence about student learning to adapt teaching to better meet student needs. His current interests focus on how school-based teacher learning communities can be used to create effective systems of teacher professional development at scale. Dylan is an Academician of the Academy of Social Sciences and a Fellow of the Royal Society for the encouragement of Arts, Manufacture and Commerce. Until 2010, Dylan was the Deputy Director of the Institute of Education, University of London. Previously, he had held posts of Senior Research Director at ETS and Assistant Principal of King's College London. His research on AfL is the most cited work in the journal *Assessment in Education: Principles, Policy and Practice* and he is internationally renowned for his empirical work and its impact upon teaching and learning.

Selection of relevant publications

- Wiliam, D (2011) *Embedded Formative Assessment*. Bloomington, IN: Solution Tree.
- Wiliam, D (2011) What is assessment for learning? *Studies in Educational Evaluation*, 37 (1), 2–14.
- Black, P & Wiliam, D (2009) Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21, 5–31.
- Black, P & Dylan W (1998) Inside the Black Box: Raising Standards Through Classroom Assessment, *Phi Delta Kappan*, October, 139–148.

© University of Oxford
Oxford University Centre for Educational Assessment
http://oucea.education.ox.ac.uk/