Assessment and learning: fields apart?

Jo-Anne Baird,[a] David Andrich,[b] Therese Hopfenbeck[a] and Gordon Stobart[c]

[a]Oxford University Centre for Educational Assessment, Department of Education, 15 Norham Gardens, Oxford, OX2 6PY, UK.

[b]Graduate School of Education, The University of Western Australia, 36 Stirling Highway, Crawley, WA 6009, Australia.

[c]University College London Institute of Education, 20 Bedford Way, London, WC1H OAL, UK.

ABSTRACT

Educational assessments define what aspects of learning will formally be given credit and therefore have a huge impact upon teaching and learning. Although the impact of high-stakes national and international assessments on teaching and learning is considered in the literature, remarkably, there is little research on the connection between theories of learning and educational assessments. Given the voluminous assessment that takes place annually in systematic ways in most many nations, it is surprising that more has not been gained from these assessments in the development of theories of learning and *vice versa*. In this article we consider both theories of learning and assessment and draw the main message of the article, that if assessments are to serve the goals of education, then theories of learning and assessment should be developing more closely with each other. We consider fundamental aspects of assessment theory, such as constructs, unidimensionality, invariance and quantifiability, and in doing so, we distinguish between educational and psychological assessment. Second, we show how less traditionally considered cases of a) international assessments and b) *Assessment for Learning* affect student learning. Through these cases we illustrate the otherwise somewhat theoretical discussion in the article. We argue that if assessment is to serve the learning goals of education, then this discussion on the relationship between assessment and learning should be developed further and be at the forefront of high-stakes, large-scale educational assessments.

INTRODUCTION

Assessment plays a central role in education. Assessments are used to investigate what people *know* and *can do* and to make decisions regarding whether they have learned what was expected. Although assessments necessarily generate observable performances from students, the concern is not merely for the performance, rather the performance is used as a warrant for inference to *competence*; although all there is to go on is performance. Even if assessment questions require people to be able to name the phases of the moon, there is usually a broader notion that this indicates something deeper about understanding of the lunar phases. In assessment nomenclature, the

subject matter, or domain, being assessed is termed the construct. Learning implies improved proficiency in the construct variable, such as reading comprehension or numeracy. We return to the meaning of 'construct' more fully below.

Our varying epistemological notions of what counts as learning frame our views about the psychological processes of learning and which pedagogical processes promote learning. Further still, these views affect what we view as appropriate assessment arrangements. Ideas about what constitutes progress in learning come from a variety of sources, but they are partly shaped by our theories of learning, whether those are formal, scientific theories or practice-based theories of action. Figure 1 shows an idealised relationship between learning and assessment theories in designing assessments. A rationalist might anticipate that theories of learning influence assessment theory and assessment design: at a minimum, learning theories shape our notions of what aspects of performance need to be included. Assessment design would both reflect and influence the outcomes of learning and therefore, theories of learning and assessment would ideally benefit from the information generated by each. Moreover, assessment theory and learning theory might also be expected to have reciprocal effects upon each other. A significant question for this field is the extent to which theories of learning have influenced assessment theory or assessment design. Likewise, empirical findings from assessments might also have been expected to influence theories of learning, so the extent to which this is evidenced in the field might be questioned. For some, the need for a correspondence between assessment and learning theories is obvious and the foregoing will seem superfluous. For others, assessment is entirely separate from learning. In this article, we seek to establish that not only does there need to be a correspondence between learning and assessment theories, but that it should be stronger than it has been to date.

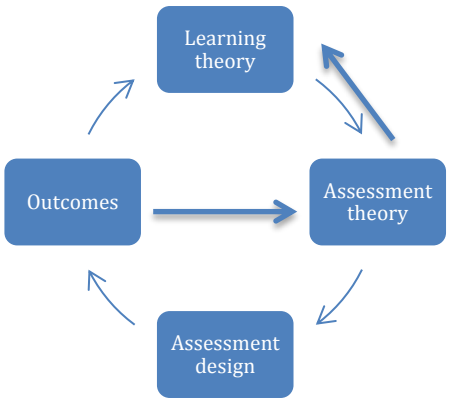

Figure 1 Idealised relationships between theory and assessment design

Learning theory foundations for assessment are not always explicit in the literature, though there are some exceptions. Some authors have taken on the difficult task of unpicking the relationships between predominant theories of learning and developments in assessment design (Shepard, 2000; Elwood, 2006; James, 2006). In this paper we refer to three main strands of learning theory: behaviourist, cognitive and socio-constructivist. James (2006) argued that controlled testing environments and multiple choice formats which focused upon outcomes are best considered to be aligned with behaviourist theories of learning (e.g. Watson, 1930; Skinner, 1953; Bandura

1969), in which there is no interest in the thought processes, only in the performance (see also Black, 1999, 120).

Cognitive theories of learning (e.g. Neisser, 1967; Sternberg, 1981) came to the fore in the 1960s and are the dominant paradigm for assessment systems currently (Haladyna & Rodriguez, 2013, 29). Nowadays, the computer model of mind metaphor that underpinned cognitive theories (e.g. Boden, 1988; Searle, 1990) has been supplanted by connectionist models and neuroscience (see Barsalou, 1992, 8-12).[1] Our metaphorical models of mind have progressively reflected contemporary technology; steam engines (e.g. Freudian psychology – see Levine, 2000), input-output processes of factories (e.g. Watson, 1930), complex systems (e.g. Vygotsky, 1978; Engstrom, 1987), computers (e.g. Anderson, 1990) and networks (e.g. Rumelhart, 1998). Metaphors of mind influence not only our theories, but also our practices. These metaphors are themselves shaped by the tools at our disposal, including statistical models (Gigerenzer & Goldstein, 1996; Gigerenzer, 1991).

James (2006) pointed out that with increasing emphasis upon cognitive theories of learning in the 70s and 80s (e.g. Simon, 1979), there was a shift in assessment design towards partial credit for appropriate methods used by students, even if the right solution to a problem was not achieved. Additionally, the requirement for assessments to test higher order thinking skills to a greater extent accompanied the transition to cognitive views of learning.

Because the tradition can be traced back to William James (1890), socio-constructivist theories of learning (e.g. Vygotsky, 1978) have essentially provided a competing paradigm to behaviourist and cognitive theories throughout the 20th century. In this approach, learning is jointly created by the learner and their social environment. Learners create new knowledge. The influence of these theories is seen in the authentic assessment movement, in which assessments are more closely tied to the learning environment. Additionally, classroom assessment practices, such as formative assessment (Scriven, 1967) are often claimed to be closely connected with socio-constructivist learning theories (e.g. Black & Wiliam, 1998a, 1998b; Torrance & Pryor, 1998; James, 2006; Shepard, 2000). Assessment formats such as portfolios, peer assessment, reflective diaries and so on have been justified on the grounds of sociocultural theory. We return to the question of whether formative assessment, or more specifically 'Assessment for Learning' is underpinned by socio-constructivist learning theories below. For now, we note that tracing the connections between learning theories and assessment design is difficult (James, 2006).

Despite the apparent lack of a solid relationship between learning theory and assessment practice, strong relationships between assessment and teaching and learning practices are claimed in the literature. The impact of assessment upon teaching and learning has been termed 'washback' or 'backwash' (Alderson & Wall, 1993). Some studies have found no negative effects of testing upon schools (Kellaghan, Madaus and Airasian, 1982), but others have found that the taught curriculum was narrowed to the material that was anticipated on the test (Au, 2007; Madaus, Russell & Higgens, 2009). Further, test washback has been found to result in more superficial learning of

[1] often what was previously termed cognitive experimental psychology

disconnected knowledge, rather than to a broad and deep understanding of subjects (Darling-Hammond & Rustique-Forrester, 2005; Daly et al., 2012). It is now accepted that the high-stakes nature of testing drives teachers and learners to change their behaviours in this way (Stobart and Eggen, 2012). High-stakes tests were defined by Madaus (1988, 29) as those that have a direct link to rewards or sanctions for students, their teachers or their institutions.

In this article, we take two disparate areas of the assessment research literature as contrasting cases (Yin, 2009). Despite their dissimilarity, they challenge the notion that washback occurs only for what would normally be considered high stakes tests in which the stakes are not direct for students, their teachers or institutions. Even though it is not direct, we trace the route by which washback occurs nonetheless. Our choice of cases reflect areas which have been distinctive, rising trends this century: A) international tests and B) Assessment for Learning (AfL). We summarise the cases briefly here before discussing fundamental issues for the field and then we return to the cases more fully to illustrate the issues.

Case A - International tests, such as the Third International Maths and Science Study (TIMSS) or the Programme for International Student Achievement (PISA), are assessments of education systems, at the level of countries or jurisdictions. Because the systems are to be compared quantitatively, assessments on multiple items are summarized with a single number for a construct or sub-constructs of assessment. Accordingly, these assessments are based upon psychometric traditions in which the idea of measurement, by analogy to physical measurement, is entertained. Standardised administration across jurisdictions is essential to the effective functioning of international tests. Assessments are constructed through a complex, formal process involving designing questions in one language (usually English or French) and translating them to other languages.

Case B – Assessment for learning (AfL), which is designed to help the teacher to diagnose where students are in their learning, so that the student can be helped to advance, are at the classroom level. Formal statistical techniques are not ubiquitous in AfL. Interaction between the assessor and the learner are key to it. Assessments are often teacher-designed and may be part of classroom dialogue. At one level, classroom-based AfL simply appears less professional than international tests. Equally, some would argue that it is only through AfL techniques that students' understanding can be gauged well.

We now turn to some assessment fundamentals to explain how the variable representing learning (the construct) is conceptualised through the lens of assessment; the construction of this variable, the fact that correlation between questions is considered to be fundamental, the tension in assuming that it is a single, unidimensional variable, problems in maintaining the meaning of the variable across contexts (invariance) and the degree to which learning can be quantified, if at all. These fundamentals matter because our assessment tools constrain and shape the metaphors of learning with which we operate.

EDUCATIONAL ASSESSMENT, PSYCHOMETRICS AND LEARNING

Educational assessment of learning is normative; *it is intended to affect the attribute being assessed*. Educational assessment is part of Foucault's (1975, 308) power of normalisation in modern society, to which individuals subject themselves. He wrote,

> The judges of normality are everywhere. We are in the society of the teacher-judge, the doctor-judge, the educator judge, the 'social worker'-judge; it is on them that the universal reign of the normative is based; and each individual, wherever he may find himself, subjects to it his body, this gestures, his behavior, his aptitudes, his achievement. (1975, 304)

Foucault's argument extends to medical and psychological assessments, which can be part of normative processes and structures. The motivational structures that are set out by assessments need to be carefully designed, lest they motivate the wrong behaviours. Not only does washback occur under certain conditions, it is *intentional* and therefore it should be recognised overtly as part of the assessment design process. Of course, there are also unintended washback effects. But the point is that psychological and educational assessment constructs are different in this regard; you are not *supposed* to improve your IQ, self-efficacy or personality in preparation for being tested.[2] The theoretical underpinnings and relationship with the act of testing differ between educational and psychological tests. Learners are expected to be active in relation to educational attributes in advance of testing (Elwood and Murphy, 2015).

One function of educational assessment, then, is as a communicative device, setting out what the curriculum designers want students to know and be able to do. Wiliam (2010) argued that assessment instruments operationalise the construct, which would then explain why assessment has such a powerful effect upon classroom practices. It is through a range of assessment (-related) artefacts that the education community understands the rules for ascribing and certifying learning.

Educational assessment constructs

'Constructs' were imported from logic and mathematics, first discussed by Pearson (1892) and were adopted by psychologists (Michell, 2013, 15). An example of a measured psychological construct is intelligence. The conceptual leap made by Pearson was that mental attributes could be measured (Goldstein, 2011), thus bolstering psychology's methods as scientific and moving away from experiential methods, with all of the subjectivity entailed. As the field began with intelligence testing, educational assessment and psychological assessment were not seen as distinct, though there are differences in theory and practice. Notwithstanding their differences, although debates about educational assessment constructs have not always used the language of psychological constructs, they have addressed many of the same issues; such as the socially constructed nature of constructs (McNamara, 2001; Elwood and Murphy, 2015), unidimensionality (Goldstein, 1979) and invariance (e.g. Stobart, Elwood, & Quinlan, 1992). Cronbach and Meehl (1955, 283) defined a psychological construct as follows,

[2] In fact there are retesting effects in IQ (Hedges, 1987), but that is a problem for the conceptualisation of the construct, whereas in education, revising and resitting are not problematical for the construct itself.

> A construct is some postulated attribute of people, assumed to be reflected in test performance.

They go on to specify that the construct to be measured comprises a universe of content, from which items that are domain-relevant need to be sampled and designed and to develop from these a test that is representative of the domain and provides an adequate sample of the domain. Psychometricians have focused their attention on whether these and other technical requirements of tests are met. Wiliam (2010) argued that the debate should shift from the technical features of assessment (how well we are assessing) to discussion about what we are assessing. Given the widespread nature of educational assessment, there is remarkably little reflection on the constructs underpinning them. Sharper conceptualisation of the construct of interest might in turn produce better instruments, with higher validity. Thus, better conceptualisation of the construct should also impact upon test validation processes. However, test validation theory (Newton & Shaw, 2014) appears to have outstripped practice, with little work being published on the validation of educational assessments (Wolming & Wikstrom, 2010). It follows that the field is riddled with *assumptions* that the attributes test designers set out to assess *are* being assessed, and that those attributes are themselves under-defined.

Constructs reflect the domain of interest and indicate progress in the domain. As such, constructs are a basis upon which scoring criteria can be developed. Proficiency statements, grade criteria or level descriptors act as depictions of particular levels of proficiency on the construct. Three distinct approaches to developing constructs are used in practice, although oftentimes they are blended:

1. Theory-based: constructs are formulated on the basis of a theory, such as Piagetian notions of the development of scientific ideas
2. Empirically-driven: constructs are devised from the results of students on previous tests
3. Subject-matter expert-devised: constructs represent disciplinary experts' views on what counts as progression in the subject

Much practice is based upon historical precedence. Missing from the above list is the political sphere, which influences the construction of assessments in many settings. Therefore, we add a fourth method,

4. Policy-driven: constructs are developed to signal to educators what politicians or policy-makers view as important

Production of educational assessment constructs differs from psychological constructs in that the theory underpinning them is often weaker, the processes can be less formal, may not be documented and politics does not usually influence psychological constructs so directly.[3] Constructs are socially constructed, but they are a social reality, with scores on educational assessments having significant implications for life chances. As variables, constructs are not observed, they are theoretical (Hood, 2008). Educational assessment constructs are not 'out there' waiting to be discovered: the realist position. Scientific

[3] Though this is not absolute. Witness, for example, the Dangerous and Severe Personality Disorder (Duggan, 2011).

realism usually entails three sets of beliefs: that theories can in principle be true or false, that the objects included in scientific theories exist independently of our theories, thoughts or language and a causal model is usually associated with a realist position. (Devitt, 1997; Borsboom, 2005). In response to the realist position, Borsboom (2005, p43) argued that he was not a walking set of constructs. From the socially constructed position, Borsboom *is* a walking set of constructs; and at that, he comprises some constructs yet to be projected upon him. Further, like the rest of us, Borsboom's identity has been socially constructed to an extent by his performances on tests (Hanson, 1994; Hacking, 2007).

Students are not born with educational proficiencies. The curriculum itself arises from the broader context of society and culture, including values. In addition, there is a more or less explicit understanding that learning in successive years of schooling builds on the learning of previous years – proficiency is accumulated. Assessments are constructed to assess the learning that was expected to have taken place. Not only this, but test features may reflect and amplify aspects of the culture in which they arise.

Assessment by association

Neuroscience offers the tantalising prospect that with the advance of technology, mental assessment could be a direct manifestation of thought, ability, personality and so on. However, this reductionist approach to mental testing does not hold enormous potential for educational assessment - certainly not at the current time and perhaps not ever. Findings about brain structure and processes have yet even to be worked up to education hypotheses before being subjected to testing (Della Salla & Anderson, 2012).

Presently, mental testing relies upon constructs being represented indirectly and empirically in relatively high correlations among items. It was anticipated that the associations would be explained theoretically once the psychological laws determining scores were uncovered (Thurstone, 1959; Michell, 1997). If correlation is all there is to underpin assessment, there are grave problems because conceptually disconnected variables correlate. For example, chocolate consumption per capita and national cognitive function correlate well (0.79: Messerli, 2012), yet it remains a possibility that no amount of chocolate-eating will produce a Nobel prize winner.

K. Pearson, who invented the statistical technique of correlation, recognised it as a scientific breakthrough that permitted research on a wider array of phenomena, for which the causes might be non-obvious and multiple (see Aldrich, 1995). Of course, he also understood that there is a range of relationships between variables that result in correlation, not only causal ones. Without causal relationships underlying the correlations between items in our assessments there would be aggregated, potentially unrelated, uninterpretable variables. For this reason, a 'nomological network' to underpin assessment was proposed (Cronbach & Meehl, 1955), which amounts to theoretical explanations for the associations between variables; observed and unobserved. The meaning of numbers generated by assessment models is generated not by the numbers themselves, but by the people generating the link between the numbers and substantive theory (Markus and Borsboom, 2013, 45). The act of assessing a construct impacts upon how the construct is perceived; it is an agenda-setting act (Maul, 2013a). Nothing about a set of numbers in themselves tells us what they measure

(Maraun, 1998). To understand what the numbers mean, factors beyond test theory must be appealed to.

However, the degree of correlation between aspects of an assessment is not to be totally disregarded. If questions correlated perfectly, there would be little point in administering an entire test. Asking a science student about the chemical reaction caused by an enzyme would tell you everything you needed to know about their understanding of science generally and you would not need to test what they knew about DNA, for example. Of course, items do not correlate perfectly. Different questions are necessary not just to reduce error through replication of measurement; they are needed to sample the domain. Educational assessment domains are not homogeneous. Test specifications typically require stratified sampling from the domain in educational assessment. Imperfect correlations are therefore at least implicitly anticipated, with consequent tensions about the extent to which they should be tolerated, and their meaning. Some formal models posit a single underlying dimension that causes scores and statistical tests of unidimensionality have been developed; we turn to that topic next. To summarise the above, assessment models have an underlying assumption that aspects of learning in a particular domain are correlated, but it can be seen that as an absolute, this assumption breaks down in theory and in practice.

Unidimensionality is relative

Whether a variable is unidimensional is typically phrased in an absolute way (e.g. Kreiner and Christensen, 2014), which in principle makes the question unanswerable, and therefore gives scope for endless debate. There seem to be four main potential traps in conducting tests of model fit. First, no model can fit any dataset perfectly, and little is said *a-priori* about what might be a satisfactory level of power to reject the hypothesis that the data fit the model. The greatest factor that governs the power of test fit is the sample size. In general there is not an a-priori specification of the degree of precision that may be needed for a particular purpose. A related factor that affects power is the relative alignment of the locations of the persons and items.

Second, the statistical tests of significance that are used to reject the hypothesis of model fit are referenced to perfection. Most significance tests are conducted to establish that the systematic factors are greater than chance, and therefore those that are checks on perfection can raise concerns that are unjustifiable.

Third, multiple statistical tests are available to understand the unidimensionality of a single data set – item fit statistics can be identified, there may be many items, each item might have a differential item functioning index, and so on. None are necessary *and* sufficient to conclude fit, and many of the statistics are approximations to null distributions, such as the Chi square distribution. These tests of fit need to be used in conjunction with each other, with professional expertise, rather than mechanistically and independently. Unfortunately it is easier to use them mechanistically.

Fourth, no item can fit a model alone – the fit is a test as to whether the items operate together in such a way that the responses to the multiple items can be summarized by a single parameter for a particular set of students. It is possible that the misfit of some item or collection of items would disappear, not just if those items were taken out of the response matrix, but if other items were taken out. Professional judgement is required

to decide whether the models fit and the degree to which a test is unidimensional. Both errors - claiming fit when there is little power, or that there is misfit when the power of the fit is absurdly high - can be made readily. Nevertheless, there is no reason to be paralysed by these exigencies – they are simply practical problems for the field.

It is generally assumed that proficiency governs students' responses to all items and that the construct is a causal one. However, sometimes items are selected to form an index – that is, the causal connection is from the item to the proficiency rather than from the proficiency to the item (Stenner, et al, 2008; 2009; Tesio, 2014a, 214b; Andrich, 2014). Essentially, in this case the items define the construct. Andrich provides the example of a test in physics composed of items on the topics of heat, light, sound electricity and magnetism, and mechanics. Because of the curriculum design, these five topics are taught distinctly within the subject of physics – they define the subject construct within the particular frame of reference. From a construct perspective, Andrich coined the terms relatively *thin* for each of the topics and relatively *thick* for the subject of physics. Subjects have also been described as composite variables (Maul, 2013b). From a statistical fit point of view, responses from a sample of well-aligned physics students to items within each of the thin variables, designed to be more or less difficult, may fit a unidimensional model. Moreover, if responses to one item from each of the five domains were chosen for the set of responses, they may fit the same unidimensional model. On the other hand, taking all items from the five topics together may not fit a unidimensional model. The reason for this is that items within each topic are likely to show a higher correlation among themselves than items from different topics. From a statistical point of view, it may appear that the response matrix is not unidimensional – there is local dependence within subsets of items which may be seen as dimensional dependence as described above. It is evident that this apparent multidimensionality is in part constructed by defining the composite variable called physics in terms of five topics, and in part produced by having more than one item assessing each topic.

Taking several of the above arguments together, we are arguing that educational assessment constructs are at once causal *and* index variables. The extent of correlation between items and therefore the construct's unidimensionality is caused in large part by curriculum design. This throws up the unhappy prospect that constructs might go too far down the index route, being composed of unconnected items forming a gibberish test. Equally, this state of affairs is uncomfortable in that statistical techniques can be depended upon less to answer questions about test quality. We are not arguing that anything goes. Rather, we are pointing out that this is a better description of the fundamental state of the art in educational assessment.

Invariance – now you see it, now you don't

Educational achievement data are integrated into the fabric of most societies (Foucault, 1975). The purposes to which data are put depend upon certain properties. If the data do not have those properties, the use of them in these ways is unwarranted, invalid and unfair. Newton (2007) listed 18 ways in which educational assessment results are used. Let us consider three of them. To compare students' results across schools, exam results need to give equivalent credit to, for example, mathematics attainment that does not depend upon some irrelevant characteristic, such as which examiner judged the work. To select students for Higher Education who sat their examinations in different years, the results need to be comparable over time in terms of mathematics attainment and not

dependent upon the relative difficulty of the questions set in a particular year. Should we wish to compare schools in terms of their students' test scores, aggregated over a range of examinations taken at age 16, the scores must not depend upon anything other than the attainment of the students, such as the relative difficulty across subjects. Invariance means that the measurements should not change due to the conditions of measurement, such as the examiner conducting the scoring or the year in which the test was taken. These requirements for invariance arise from the uses to which put test scores are put and not from any particular testing or statistical model. Discussing attitude scales, Thurstone put it as follows,

> If the scale is to be regarded as valid, the scale values of the statements should not be affected by the opinions of the people who help to construct it. … the scaling method must stand such a test before it can be accepted as more than a description of the people who construct the scale. (Thurstone, 1959, p228)

Obstinately, real-world educational assessment data do not always fit these requirements (Mislevy, 1997). Scores depend upon the particular items presented, the design of the test, the background characteristics of the students and so on. Our comparisons, then, are based upon scores that do not have the absolute property of invariance that is sought. Understandably, this leads to academic controversy, as well as public outcry.

Yet examination scores *are* used and have not all fallen into complete disrepute. There are a number of reasons for this. First, they are better than alternative systems on the grounds of invariance (e.g. nepotism). Second, invariance is not absolute, there are degrees of invariance, so the issue becomes the extent to which they are tolerable, rather than whether invariance exists. This is ultimately a political question, with examination systems serving to legitimate the replication of existing societal structures. Third, educational assessment is not static; it is subject to a great deal of reform (Berry and Adamson, 2011). This causes a lack of invariance because the attributes of interest and the instruments used to assess them shift. Continuous reforms are in part a product of the political nature of educational assessment as a tool. So some lack of invariance is caused by political tensions regarding equity of the allocation of societal resources. In turn, the tensions explain the, sometimes sanguine, tolerance of invariance, to the extent that people understand the deep-seated role of power in the design and operation of assessment systems. Equally, the political tensions explain why assessment is sometimes reviled.

Invariance is not only a product of the political, however. Learning is idiosyncratically individualistic, context-dependent and socially produced, all of which are recipes for invariance. Thus, the question that arises repeatedly is the extent to which examination scores are a product of the measuring instrument or of the underlying attributes of interest. To illustrate the issue: intelligence tests were *created* with the *belief* that intelligence was normally distributed in the population. Questions were designed and selected so that the test results fitted this pattern. Humans did not discover the normal distribution of intelligence, any more than people discovered that women were smarter

than men.[4] Having done so, the normal distribution proved very useful because centuries had been spent understanding its mathematical properties, as a basis for statistical analyses in other fields, beginning with astronomy (Porter, 1986). Thus, the distribution of intelligence was a matter of utility, not fact. More deeply, it was a product of an underlying belief in the special nature of the mathematics of the normal distribution because it was encountered in physical, biological and social phenomena. Awe at the capacity of mathematics to explain our world and universe has a long tradition, with Plato being a key protagonist.

Quantifiability – what if the truth isn't out there?

The above begs the question about the quantifiability of educational attainment in the first place – a question that has previously been raised in the literature on psychological constructs (Michell, 1997a, 1997b, 2005, 2008b, 2013). Perhaps attainment performances are not matters that can be turned into measurements. Perhaps the relationship between attainment and assessment scores is non-linear and far more complex than its treatment recognises (Borsboom, 2008; Cronbach and Meehl, 1955). For example, many British universities use a 0-100 score system, in which 60 to 70 is an upper second class degree and over 70 is a first class degree. What if the attainment required to move from 66 to 71 is greater than to move from 60 to 65? Given that overall degree results are calculated from module scores, this matters because students would have to do *far* more to get a first class degree than our numerical system implies, compensating for lower scores on one module by much higher (underlying) attainment elsewhere.

Michell (2008a) laid out ten requirements for attributes of assessment to be considered quantitative – they are the conditions for real numbers. Indeed, Michell (1997a, 1997b, 2005, 2008b, 2013) has repeatedly questioned whether psychological constructs can conform to these measurement principles, as applied in the physical sciences. Kane (2008) responded that he did not think the educational assessments would meet these strictures either:

> Taking achievement in chemistry as an example, different people, *a* and *b*, would typically have different patterns of competence. Person *a* might be good at solving numerical problems but perform badly in the lab, and person *b* might show the opposite pattern. Which person is higher in overall achievement in chemistry? Given an area of achievement that is broadly defined, we are likely to have, at best, a partial ordering, unless we arbitrarily decide that some patterns are better than others.

Since educational assessments are so often used to select the top performers, who have typically had to answer most of the questions correctly, any lack of linearity might have troubled people less than had there been more interest in cut-scores lower down the scale. Yes, there will have been some students misclassified, but unless the assessment

[4] Following analyses of sex differences in IQ scores, Weschler wrote that "we have a 'sneaking suspicion' that the female of the species is not only more deadly, but also more intelligent than the male." (1944, p107). Many IQ tests are designed to eradicate sex differences, though research has reached varying conclusions.

was hugely error-prone, the students who scored highly were good at least, at all of the sub-components of the examination.

From the discussions above it can be seen that some indications arise in educational attainment data that lead us to question the strict quantifiability of attainment. Taking a step back, it would be absurd to think even of physical properties as quantities in themselves. Mass, energy and electricity are not quantities, but they are quantifiable. They are quantifiable because quantitative models have good utility – they are predictive, albeit imperfectly, due to the vagaries of the real world. The question, then, *is not* whether educational attainment exists as quantity in people, but whether attempts at *quantifying* are useful and predictive.

Our argument is not solely functionalist or operationalist (Borsboom, 2005), as we are interested in the nature of educational attainment itself. Neither are we proposing that "only the features causing differences in performance are quantitative" (Michell, 2008c). Instead, at this stage in the field of educational assessment, the nature of educational attainment itself, processes causing it and the relationship between it and assessment results are all matters for serious research and may, or may not ultimately usefully be construed as having quantitative structures. The extent to which quantification is useful depends upon the subject matter being assessed, the nature of the assessment, how the scores are used, as well as who is judging their utility.

Given that measurement in natural science provides a prototype for the social sciences, it is relevant to consider the function of measurement in the processes of scientific discovery. The standard proposition is that measurement leads to scientific theories, but Kuhn's (1961, 193) analysis suggests otherwise:

> In textbooks the numbers that result from measurements usually appear as the archetypes of the 'irreducible and stubborn facts' to which the scientist must, by struggle, make his theories conform. But in scientific practice, as seen through the journal literature, the scientist often seems rather to be struggling with facts, trying to force them to conformity with a theory he does not doubt.

If discovering theories is not the role of measurement (Maraun, 1988), as portrayed in text books, it seems necessary to understand the role it has in the advancement of science. Kuhn (1961, 180) gives the answer:

> To the extent that measurement and quantitative technique play an especially significant role in scientific discovery, they do so precisely because, by displaying serious anomaly, they tell scientists when and where to look for a new phenomenon. To the nature of that phenomenon, they usually provide no clues.

It may be interpreted that Kuhn implies simply that measurements are obtained using existing instruments and that from predicted relationships possible anomalies are identified. This interpretation seems too simple. We take it that the theory directs the measurement itself, including instrumentation, and that from a theoretically governed attempt at measurement, anomalies are discovered. This interpretation implies that the understanding that places variables in a quantitative theoretical framework and the

understanding that permits the construction of instruments to measure the variables generally go hand in hand. This joint relationship is evident in the development of the now familiar measurement of temperature, where for a substantial period scientists were measuring temperature to some useful degree of precision, without understanding temperature in the way they understand it now (Choppin, 1985; Webb, 2013). The close relationship between theory and measurement extends beyond temperature to other measurements in physics (Humphry, 2013).

THE PLACE OF THEORY IN EDUCATIONAL ASSESSMENT

Two distinct sources of theory are relevant - theories of assessment[5] and substantive theories, in this case of learning and attainment. Given the postmodern position outlined above, it is clear that the relationship between theories of measurement and substantive theory ought to be compatible, or our instruments will be confined to portray only aspects of the phenomena of interest. Indeed, this seems to us to be the current state of affairs.

Assessment and learning theories seem to be fields apart, with scope for far greater connection. Given the volume of educational assessments that have been conducted, the paucity of contribution to knowledge about learning from psychometrics is striking (Blinkhorn, 1997), though some researchers (e.g. Wilson, 2005, 2006, 2009) are trying to remedy this through evidence-centred design (Mislevy et al, 2003). Of course, it is possible to see it as the job of substantive theory to indicate the nature of the attributes and by implication the assessment instrument, in which case the fault lies with weak substantive theory (Borsboom, 2006). Models of learning that underpin test design are often either not referred to, or remain in 'puberty, infancy or even at the foetal stage' (Sitjtsma, 2006, 453; see also Mauran, 1988).

More broadly there are concerns about whether the field of education is advancing, by building knowledge cumulatively (see Oancea, 2005; James, 2012). Education has been cast as an immature field (Wieman, 2014) that lacks the underlying causal models that took centuries to build in the physical sciences (Mari, Maul, Irribarra and Wilson, 2013). Gould (1996) wrote that the field was suffering from 'physics envy'. The physical sciences are viewed idealistically in this literature, though, as some of the difficulties faced in education, such as consistency of findings, are found in the physical sciences too (Hedges, 1987). As Kane (2008) wrote, developing science is messy.

Whether mental attributes are suitable for quantitative treatment has been questioned generally (Michell, 1997a, 2008b, 2013; Hood, 2008). Our position is that quantification

[5] Rightly, it has been pointed out that mathematical formulations are not theory by themselves (Goldstein and Wood, 1989; Laming, 1997, 390). By measurement theory, we mean going beyond a mathematical formulation to the justification for the model and its application (Andrich, 2004; Humphry, 2013).

is a valid way to proceed, though the utility of quantification needs better research and justification. However, we are not proposing the operationalism that has been alleged (Borsboom, 2005), because it is possible to maintain a pragmatic, utilitarian perspective simultaneously with a keen interest in substantive theory. This is neopragmatic, postmodern test theory (Mislevy, 1997). Though advances in the theory of measurement should provide a more refined conceptual basis for the assessment field, as stated above, we do not see them as an end in themselves for the understanding of the phenomena of interest: educational attainment.

As the assessment instrument has a role in shaping the observed substantive phenomena (Moss, Pullin, Gee, & Haertel, 2007; Maul, 2013a), tensions arise between assessment theory and substantive theory when the data do not fit the assessment model (Choppin, 1985; Goldstein and Wood, 1989; Goldstein, 1979). Questions then follow regarding whether the data (instrument) need to be improved, or if the model should be revised. Further, the model or assessment procedure can come with its own internal logic and apparatus, which can take supremacy over substantive theory considerations. Indeed, some of these tensions have arisen due to a clash of measurement theory paradigms – that between assessment and psychometrics.

Assessment has been distinguished from psychometrics as follows (Gipps, 1994): assessment

- does not view 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 (rather than 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.

Essentially, Gipps argued that assessment was based upon more authentic tasks and justified the philosophical positioning of assessments with reference to postmodernism (Gipps, 1994; Lewy, 1996). Psychometrics is still generally grounded in modernist practices (Michell, 1997; Mislevy, 1997; Borsboom, 2005). Here, the philosophical positions, learning theories and measurement model paradigms are in opposition. Neither this 'assessment' approach, nor psychometrics are neutral tools in the quest to understand learning. Each brings sets of assumptions regarding test formats, processes and procedures that educationalists might find unhelpful. Our very definition of what it means to have learned will be shaped by the instruments and apparatus that people bring to the task. Moss et al., (2005, 70) wrote that,

> … 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…

Baird and Black (2013) argued that educational purposes need to be brought to the fore to a larger extent for the field to move forward and that, due to its strictures, at times psychometrics could seem like an answer to somebody else's problems.

Notwithstanding, the technocratic[6] knowledge and power that come with psychometrics and their industry-backing mean that there are significant power struggles over what kinds of learning are to be valued.

There is widespread recognition that better theory about the substantive content is needed to motivate assessments (Maul, 2013b). For some, psychometrics is atheoretical and has caused the lack of substantive theory through rail-roading and over-promising (Goldstein and Wood, 1989). For others, substantive theorists just haven't given psychometricians enough to go on (Borsboom, 2006). It is hard to avoid the conclusion that the way forward needs to be a joint enterprise (Andrich, 2004), but it needs to be recognised that the priorities of assessment experts, educators and policy-makers are not necessarily going to align neatly (Baird and Black, 2013).

To summarise, educational attainment constructs set out what students should learn. Unlike psychological constructs, they are goals. Within an education system, they help to generate the very attributes that they assess by making transparent what students should know and be able to do. Indeed, viewed strategically, this is arguably the *main* function of educational assessment and therefore the design of our assessment systems should better reflect valuable learning. 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.

Next, we turn to the two cases to look at the assessment and learning theory paradigms within which they operate. These cases serve to illustrate the distinctive positions that can be taken on the above issues. We discuss the effects upon these assessment practices upon learning.

CASE A - INTERNATIONAL TESTS

Formal underpinnings

International large-scale assessments (ILSA) of educational constructs include those produced by the International Association for the Evaluation of Educational Achievement (IEA) and the Organisation for Economic Cooperation and Development (OECD). For the sake of brevity, we will refer mainly to the Programme for International Student Assessment (PISA). The main purpose of having international tests is to be able to compare across countries how well the students perform, entailing the assumption that it is possible to make these measurements and comparisons (Gustafsson, 2012). Thus, we assume quantifiability. Also, it is necessary for ILSA to assume invariance; that the measures do not change with context, but have the same meaning whether they were taken in China or Peru. Methodologically, a psychometric approach is used that assumes a unidimensional construct (OECD, 2014). Psychometrics is best aligned with a

[6] Technocracy is a system in which decisions are made by people with technical knowledge (see Lawn, 2008).

realist position (Borsboom, 2005), in which for example, mathematical literacy is a property of individuals that exists independently of the test.

Learning theory underpinnings

The independence of ILSA test scores from the learning environment is treated differently by OECD and IEA. Studies such as the Trends in International Mathematics and Science Study (TIMSS), generated by IEA, are based upon a *curriculum-related* approach, in which the test content is agreed as common to the participating countries' curricula. It can be argued that comparisons are then valid, as they refer to the international curriculum (Stanat and Lüdtke, 2013), though they do not tell you about the other curriculum material, which could differ markedly across countries. Thus, the invariance relates to the shared curriculum material. OECD does not attempt to construct a consensus-based test, instead using a *literacy-oriented* approach. This is a stronger claim for the realism of proficiency on the test, which assumes that the construct can be measured independently of the country's curriculum. Test items require students to apply their knowledge in questions set in more authentic contexts and to solve problems that they might not yet have experienced in school (Nardi, 2008). Literacy requires not just knowledge, but the universal cognitive skills of application, analysis, evaluation and reflection (OECD, 2001, Freebody and Freiberg, 2011). In a technology-rich world, some argue that generic skills will be more important than knowledge and that a curriculum-based approach is not fit for the future (Simon, 1996; Østerud, 2006). OECD is not the only assessment authority to have taken the turn to generic skills and there has been concern that the focus upon generic skills has undermined the knowledge component of education (Young, 2008).

If we look more carefully at the items used in these studies, it is hard to find the cognitive underpinnings of the international test constructs. Using reading literacy in PISA as an example, the original framework used the IEA Reading Literacy Study (Elley, 1992) and the International Adult Literacy Survey (IALS) and the work of a few reading researchers. The 2000 OECD report *Measuring Student Knowledge and Skills* built upon the work of Applebee et al (1987) with its focus upon readers' abilities to analyse, evaluate and extend ideas presented in texts. Reading literacy in PISA 2000 was defined as:

> The capacity to understand, use and reflect on written texts, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society. (OECD, 2000, 10)

In the fourth PISA cycle, the PISA 2009 *engagement* was added to the definition to the reading literacy, as follows:

> Reading literacy is understanding, using, reflecting on and engaging with written texts, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society. (OECD, 2009, 23)

The framework in 2009 reflected a stronger research influence by scholars such as Guthrie (2008), Guthrie and Wigfield (2000), and Guthrie, Wigfield et al (2006) on engagement and motivation to read, as well as the work on reading comprehension by Michael Pressley et al (1989 and 2006) and Dole et al (1991). The authors of the 2009

framework further claimed to be influenced by contemporary theories of reading, such as Graesser, Millis and Zwaan, (1997) and Kintsch (1998), but it not clear how these works influenced the framework.

Despite the changes in the framework and more focus upon reading research in 2009, when presenting the results from PISA 2009, the test appears to be more data-driven than theory-driven. Originally the PISA framework was measuring five aspects of reading literacy, and developed items which would measure these: 1) Broad understanding (20% of items), 2) retrieve information (20% of items, 3) developing an interpretation (30% of items) 4) reflecting upon the content of a text (15% of items) and 5) reflecting upon the form of a text (15% of items). After conducting the first PISA cycle in 2000, it was decided to report the students' results in three categories: 1) process skills, 2) knowledge and understanding and 3) context of applications (OECD, 2001). The process skills included the readers' ability to retrieve and interpret information from text with the focus upon tasks students would need to know in real life. The second dimension of reading capture the content types of text, whether it is a graph or a text in prose, students' need knowledge and understanding of it. The third and final dimension captures the context or purpose of the text, and demands of the reader to understand the text as the author intended the text to be used. Thus, it would appear that once the results of the first PISA cycle had been analysed, the data indicated that only three sub-scales could be supported. As discussed earlier, this begs the question of whether the test data are caused by the nature of the attribute – reading literacy – or by the nature of the test.

As for each PISA cycle, all domains have consulted reading experts, who are also listed in the PISA publications, but as PISA has a governing board where each country is represented by policy makers, decisions regarding which items will be used, and how the different domains are measured, are also a result of negotiations with policy makers;

> The original reading literacy framework for PISA was developed for the PISA 2000 cycle (from 1998 – 2001) through a consensus building process involving reading experts selected by the participating countries and the PISA advisory groups.
>
> (OECD 2010b)

Under such conditions, it is understandable that empirical evidence and theories from reading research are not the first priority in the development of PISA items; after all, stakeholders from more than 70 countries are involved in the process, and agreements have to be made among them. The constructs are in composed through all four processes outlined earlier: theoretical, empirical, experts and politics.

Knowledge about learning gained from international tests

International large-scale assessments are a technically rigorous approach to addressing questions about how much children have learned about a particular subject across countries. By using the sub-scales, it is also possible to compare the kinds of knowledge and skills that are not so well tackled by children in one setting compared with another. Individual items can also be instructive about the ways in which children have learned to approach problems.

One example is the PISA 2000 study, and the Mathematic item *Continent*.[7] Students were presented with a map showing the Continent Area of Antarctica, and the following questions: *Estimate the area of Antarctica using the map scale. Show your working out and explain how you made your estimate (You can draw over the map if it helps you with your estimation).* Full credit (2 points) were given to correct method and getting the correct answer, while partial credit (1) points were given to responses using the correct method but getting an incorrect or incomplete answer. The released PISA item booklet reads: "The aim is not to see if students can express well in words. The aim is to try to work out how the student arrived at his/her answer." (p20). In other words, this item has focus upon how students approach an item, what they are thinking and how they will try to solve it, there is several possible solutions possible, and the coding guide underlines the importance of looking for students' drawings on the map, to investigate whether they were thinking along the correct lines for solving the problem. It turned out that this item had been a challenging one internationally, with as much as 48% of all students internationally leaving the item blank. This particular item was challenging for Norwegian students. In fact, the other Nordic countries performed better than the Norwegian students on this item, with 18% of the Finnish students scoring 2 points, but only 7% of the Norwegian students scored full score (Lie et al., 2001, 184). The Norwegian researchers could not explain this pattern, but the results and presentations of the item generated discussion regarding students' approaches to learning, their knowledge of items and teaching instructions in mathematics. The Norwegian Research Council funded video studies in Norwegian classrooms to investigate problematic findings in PISA (Klette & Lie, 2008; Klette et al., 2015). Of course, individual items should only be used to *illustrate* general patterns because any items is a combination of language, format, cognitive demand, topic and so on, which makes it difficult to unpick which aspect of the item caused children to respond the way that they did. Individual items can be an unreliable indicator of learning.

ILSA have the capacity to inform us regarding learning, but few researchers have used the results in that way (Baird et al., 2014). Indeed, few subject-specific analyses using the content of the items have been published, though there are some exceptions, such as the special issue of *International Journal of Science Education*: Students´ Interests in Science across the World: Findings from the PISA study (Olsen et al., 2011). After 15 years and six cycles of PISA surveys, leading reading research journals such as *Reading Research Quarterly, Reading Psychology, Reading and Writing* and the *Journal of Research in Reading* have only published six articles where researchers have used data from PISA (Baird et al., 2014). So why are the data from ILSA, which are freely available to researchers, not used more systematically to investigate students' understanding and thereby improve our knowledge of their learning? We offer some reasons below.

ILSA barrier 1 – the items are not publicly available

Although the data relating to students' responses are available, the actual questions that they answered are not publicly available. Only a small proportion are released (Kreiner and Christensen, 2013). Categorisations of the question are available; for example on the question type, the level of demand, the topic addressed. Still, it is difficult to make sense of what students have learned if you cannot see what they have been asked to do. Not all researchers are restricted in this way. Those who administer the tests have access to the

[7] Released booklet and items: http://www.oecd.org/pisa/38709418.pdf

items and can gain a deeper understanding of the findings, but this level of access is not available to all. Furthermore, researchers must keep the bulk of the items secure so that they can be used to anchor the demands of the next test when it is administered. This level of obfuscation is important because without sight of the items, the data can only be interpreted through the lenses of those who constructed the test.

ILSA barrier 2 – plausible values are not transparent

Data collection for ILSA is conducted using an incomplete design, in which students do not take the same items and the difficulty of the items is calibrated using psychometric models. An overall score for each student is produced, and projections of their likely performance on items they were not exposed to is imputed (Wu, 2005). These are called 'plausible values'. Replicating their construction would enable researchers to better understand their meaning, but this requires considerable psychometric skill. Even those who do have the skill have struggled to replicate the plausible values because many 'craft' decisions need to be made that are not publicly known or well documented. Without being able to trace back the connections between plausible values, raw scores and the content of items, concrete conclusions about learning cannot be drawn from ILSAs.

ILSA barrier 3 – better data are available

Naturally, any particular ILSA can only cover a restricted range of topics, in a specific manner. Researchers might well be interested in aspects of learning that are not covered, such as the learning of practical skills in science or oral communication skills. Further, ILSA are standardised across administration cycles to maintain their standards and radical changes of the content and style would compromise that process, so they change slowly. Responsiveness to changes in theory and practice regarding curriculum, learning and assessment are therefore difficult. Notwithstanding, researchers might simply be interested in researching an area of learning with a theoretical approach that does not align with the mode of testing in ILSAs. Additionally, although questionnaires are administered with the tests, they contain only a subset of the questions that are interesting to researchers who are investigating learning. Many national datasets are available which provide better material for secondary data analyses. However, this is not true in all countries and some have uncovered issues from the PISA data that were previously not well documented. For example, in Germany, the weak performance of ethnic minority students was brought to the fore by the PISA results (Ammermueller, 2007). To use ILSA data to better understand learning, the content of the tests has to coincide with the features of learning of interest; clearly, they will not always be coincident.

ILSA barrier 4 – comparisons between countries are problematical

The main purpose of ILSAs is to compare students' performances between countries, but this area of research has proved to be problematical. Issues generated by lack of correspondence with the school curriculum were discussed earlier (Nardi, 2008). Additionally, the definition of the school-age population that are the target of the tests can causes problems for comparisons. Migrant workers' children (Baird et al, 2014) or children from poor rural families (Loveless, 2014) might not have the right to attend school in some jurisdictions. With these groups of children typically having weaker

educational outcomes, and being included in the data in some countries, interpreting between-country differences is not straightforward. Even if the population is not controversial, the sample of students tested can be controversial due to issues such as response rates (Prais, 2004, 2007; Elvers, 2010; OECD, 2010a), the grades that the students are in (Wagemaker, 2008) or the extent to which exclusions for issues such as disabilities have been applied (Hilton, 2006), or poorly reported excusion rates that exceed international standards (Rukowski and Rutkowski, 2015). The effects of motivation upon test performances might differ across countries, as there are cultural differences in national pride and how it is enacted (Eklöf, 2010; Hopfenbeck and Kjærnsli, 2016). Additionally, the tests must be translated into different languages, which can impact upon the demands of the questions (Grisay, 2003; Wu and Ercikan, 2006; Grisay and Monseur, 2007; Wiliam, 2008; Hauger and Sireci, 2008; Ercikan and Koh, 2009; Grisay et al., 2009; Le, 2009; Solano-Flores, 2009; Arffman, 2010; Elvers, 2010; Babiar, 2011; Oliveri and Ercikan, 2011; Mesic, 2012; Sandilands et al., 2013; El Masri et al., 2016). Thus, many factors mean that using the data to compare countries to draw conclusions about how well students are learning in different education systems is not straightforward.

Effects of international tests upon learning

There are at least two reasons that ILSAs might not have an impact upon students' learning, but they do. First, in the absence of firm conclusions about what students have learned from the tests, due to the barriers outlined above, it might be anticipated that the data would not lead to effects upon learning. Second, effects upon learning are typically envisaged as classroom effects; practice effects. However, it is now recognised that although ILSAs do not generally impact upon students' learning directly, they affect educational policy, which has a big impact upon what and how students are taught and learn in class. Austria, Germany, the Netherlands, New Zealand and South Africa developed reading standards aligned to the Progress in International Reading Literacy Study (PIRLS), run by the IEA (Schwippert and Lenkeit, 2012). In Japan, the national tests were changed to look more like the PISA items (Schleicher 2009) and in Norway, the PISA reading framework served as a model for the national tests in reading (Frønes et al 2012). How ILSA data are being used to draw conclusions is unclear and a number of authors have pointed out that pre-existing policies were justified on the basis of the findings (e.g. Lawn and Grek, 2012).

Conclusions on ILSAs

ILSAs are influenced by cognitive theories of learning, but the constructs are also influenced by political consensus. They are based upon a psychometrics paradigm, with assumptions of unidimensionality of the construct and a realist philosophy (Borsboom, 2005). Invariance of comparison is important in ILSAs. To summarise the above, ILSAs have not yet taught us much about learning or contributed to theories of learning, but they have had an impact upon what is learned and how it is learned through education policy. Given the lack of transparency of the data, the sparse analyses of what students have learned and the sometimes pseudo-use of ILSA data by policy-makers, there is room for improvement in the link between assessment and learning in this case. Next, we turn to the major classroom-based assessment movement of this century.

CASE B - ASSESSMENT FOR LEARNING

Formal underpinnings

Assessment for Learning, as the name indicates, seeks to link assessment directly to the learning process. As an approach it can be contrasted with Assessment of Learning, the summative use of assessment which primarily measures what has been learned (Gipps, 1994). In this case we consider how Assessment for Learning (AfL) operates with differing demands and constraints to those of standardised assessments and explore how it relates to learning theory. In common with standardised assessments, its impact on learning has to be evaluated, particularly as there is debate about overstating AfL's effect on learning. AfL focuses upon classroom practices and interactions, which can produce quantitative or qualitative information. The driver for this kind of assessment is to improve students' understanding, which can mean that the practice knowledge of the teacher is essential to knowing what the next step should be for the student (Baird, 2010). Teachers' own theories of learning are generated by all four of the methods for generating constructs; their knowledge of learning theory, empirical experience of students' learning, exposure to experts' views and policy-derived practices that inform the system about what is valued by educational structures such as inspection regimes or funding bodies. The extent to which each of these methods of deriving constructs influences individual teachers practicing AfL is of course unknown. Normative theories of learning are not essential to AfL and a unidimensional construct is not a requirement either. Formal statistical models are not necessary. Further, AfL can be practiced without signing up to a realist position.

Since AfL is practiced through teachers' and students' judgment, invariance is enacted through reliability of teachers' assessments. The focus is on the dependability of the interpretation of student performance (Stobart and Eggen, 2012). Are teachers able to identify accurately where learners are in their learning, do they have sufficient domain knowledge to understand students' progress? Similarly, do the students have sufficient grasp of the standard to be achieved (success criteria) to be able to able to assess their own and others work? Bennett (2011) claimed that in formative assessment too little attention has been paid to the *interpretation* of the evidence. The inferences made are always uncertain and 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; Bennett, 2011).

Definitions

Assessment for Learning is often used interchangeably with formative assessment (FA), a term historically associated with Scriven (1967). The conventional discussion of its origins treats AfL as evolving out of the behaviourist mastery learning associated with such as Benjamin Bloom (1956). We take a different approach here and argue that many of the key elements of Assessment for Learning, for example the role of the learners in directing and regulating their learning, draw on learning theories which pre-date or were in opposition to behaviourist formulations. As we will see below, many researchers have argued that the learning theory approach is sociocultural, which is aligned with a postmodern perspective.

The dilemma here is whether to distinguish AfL from FA (Black and Wiliam 1998, 2003, Wiliam 2011[8]; Swaffield 2011; Stiggins 2005) or to treat the terms as equivalent. According to Crossouard (2011), formative and summative assessment were reframed as 'Assessment for Learning' to avoid unwelcome jargon for teachers. For present purposes we treat Assessment for Learning as offering some distinct emphases within formative assessment approaches. While there is a common function (to directly improve learning), what distinguishes AfL is the focus on student self-regulation and autonomy and on viewing it as an informal and continuous process. This contrasts with the historical emphases in FA, particularly in the United States, on teachers using test data as feedback to adapt their instruction and on a formative assessment being seen as a product, typically a test. More recently the active student role in day-to-day assessment has received more attention (Brookhart, 2007; Cizek, 2010, Popham, 2008).

The shift in terminology, particularly outside the US, from Formative Assessment to Assessment for Learning was also intended to bring more clarity about purpose and processes and to deal with the problem of Formative Assessment becoming such an umbrella term that almost any assessment could be labelled a formative assessment (Swaffield 2011). The international take-up of AfL has led to a range of definitions that in part reflect the culture into which it is being introduced. It has, in turn, also become an umbrella term under which a variety of practices shelter. An early and widely used definition was that of the UK's Assessment Reform Group (2002):

> … 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.

The increasing focus on learners' self-regulation is found in the definition which emerged from the Third International Conference on Assessment for Learning held in Dunedin, New Zealand (Klenowski, 2009, 264):

> 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.

This definition is notable for its lack of reference to planned assessments. By contrast definitions coming from other cultures link AfL to more formal assessments. So, for example, in the US, Popham (2008, p6) emphasised the planned and episodic nature of the assessment: 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. Similar linking to more formal assessment demands can be found in more examination focused cultures such as Hong Kong (Kennedy et al. 2008; Carless 2005).

[8] Though their definition of what constitutes formative assessment is more specific than common usage : '…assessment becomes *formative assessment* when the evidence is actually used to adapt the teaching to meet student need' (1998, p. 140)

Assessment 'for' learning and assessment 'of' learning are accessible ways of presenting the different functions of assessment - the formative and the summative. There is broad agreement that this is a complex relationship in which functions overlap (Harlen and James, 1997; Black et al., 2003; Hausknecht et al., 2007). In his influential critique of formative assessment, Bennett (2011) takes issue with this oversimplification as it 'absolves summative assessment from any responsibility for supporting learning' (p7). Yet summative assessment can play an important part in learning, with the preparation for good quality tests offering a positive and powerful motivation for learning (Biggs, 1998; Kennedy et al. 2008). Taras (2009) extends this argument by claiming that summative assessment is at the heart of formative feedback since it involves a summative judgement about where the learner is.

There are dissenters from this position, with Roos and Hamilton (2005) arguing it 'is unhelpful to treat them as opposite sides of the same thing' (p.18). Swaffield (2011) and Klenowski (2009) also seek to keep summative assessment at some distance from the processes of AfL. Bennett's (2011, p7) own position is that,

> 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.

Learning theory underpinnings of AfL

AfL is an approach which needs to be adapted for each subject rather than a general framework which can be directly applied to any subject (Bennett, 2011). These practices are justified by drawing on learning theory, though this is not a specific 'AfL learning theory' but rather borrowings from more general learning theories. The risk here is that the theoretical underpinnings are often left implicit so that AfL becomes a series of classroom practices (for example, learning intentions, wait time, comment-only marking) justified and validated in terms of 'what works'.

Where there is an appeal to learning theory, this is often eclectic, drawing selectively from different traditions. So for example a central tenet of AfL is to make clear the learning objective to the learner and this could be justified by an appeal to behaviourist theory' requirement to specify what is to be mastered (Bloom 1956; Gagné, 1985); to cognitive/constructivist theory's emphasis on cues in concept formation (Bartlett, 1932; Bruner, 1960); or socio-cultural theory's emphasis on constructing the classroom learning contract (Pryor & Crossouard, 2008).

An over-simplified and debatable account of the relationship between theoretical traditions and teaching and learning practices encouraged by AfL can be found in Table 1. This also demonstrates the considerable overlap between the traditions. This is particularly so in relation to constructivist and socio-cultural overlaps, with social constructivism uncomfortably spanning the two - 'this merged middle-ground theory' which is borrowed from cognitive, constructivist and sociocultural theories which 'are sometime warring with each other' (Shepard, 2000, 6).

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

Theoretical orientation	Associated with	FA 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 instruction. Feedback to student corrective
Cognitive / Constructivist	Piaget Bartlett 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, Brookhart Cobb Sfard Assessment Reform Group (ARG) Black and Wiliam, Swaffield	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 & Pryor Pryor & Crossouard Black and Wiliam Ecclestone, Gipps, James, Allal, Perrenoud	Learner identity and changed teacher role and identity Negotiating understandings of task and quality criteria Apprenticeship model of learning, Communities of Practice 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.

Adapted from: Baird, Hopfenbeck, Newton, Stobart & Steen-Utheim (2014)

Knowledge about learning gained from AfL

In common with other areas of assessment, AfL findings have not been used systematically to influence theories of learning broadly. However, AfL is a distinctive, practice-based approach to assessment and its impact upon theories of learning is more likely to be at the practitioner level, in terms of theories of action, grounded theories, or self-theories. Teachers might well gain better pedagogical content knowledge and understand how to explain the field to their students through AfL.

Effects of AfL upon learning

The main validity argument for Assessment for Learning is consequential validity (Stobart, 2012) – how effectively it meet its purpose of improving learning. This is a contested area particularly as Black and Wiliam's early estimates (0.4 – 0.7 effect sizes), based on their influential 1998 review of the existing literature, were treated as established effect sizes in many subsequent reviews. Bennett described this as a

'mischaracterisation that has essentially become the educational equivalent of urban legend' (2011, p12). This problem is compounded by the lack of quantitative research into the impact of AfL. A systematic review found 907 relevant journal articles on AfL of which only two measured effects quantitatively and only used a randomised control design; the majority being small-scale case studies and action research (Baird et al., 2014). While most studies found positive effects, these were generally reported in terms of teacher and student perceptions of change. Kingston and Nash (2011, 2015) found a much smaller effect size (0.20) than had previously been reported for AfL, but Briggs et al. (2012) argued that their review excluded too many articles. Thus, the answer regarding the impact of AfL upon learning is still disputed.

Evaluating the impact of Assessment for Learning is further compounded by the breadth of its definition. Dunn and Mulvenon (2009, p2) concluded that the 'vagueness of the constituitive and operational definitions directly contributed to the weaknesses found in the related research'. Bennett (2011) called for a 'theory of action' to identify the characteristics and components of the concept and how these work together to provide the desired outcome. So, for example, the systematic evidence on the impact of feedback is a more focused area for evaluation. Further, Sadler (2007) has called for 'appropriate conceptual foundations for both pedagogy and assessment followed by practices that are consistent with them.' making the claim that many of the terms in the discourse of assessment is used loosely, such as learning, criteria and standards. Wiliam (2011) and Black (2015, 2016) have clarified many of the concepts on formative assessment originally suggested in their articles from 1998 (Black and Wiliam, 2009). In his review chapter on formative assessments definitions from 2010, Wiliam further suggested that a definition on formative assessment ' in terms of the function that assessment evidence fulfills: specifically the extent to which assessment supports and improves instructional decisions' (Wiliam, 2010: 22), a message also found in the 'Formative Assessment as a cyclic process model' suggested by Harlen (2006, 2016), where one of the steps is to use feedback from students to adjust teaching. Still, there the breadth of its definitions could be seen as a challenge for the field where different definitions are used across the research field.

Conclusions on AfL

Assessment for Learning claims a direct link between assessment and learning. Assessment is largely defined in terms of informal classroom processes which identify where learners are and the feedback that can help reach the desired standard. Periodic summative assessment may be used formatively to play a role in this.

The theoretical basis of AfL is eclectic and often implicit. Different cultural and learning theory traditions lead to differing focuses, for example the American behaviourist tradition placing more emphasis on testing and the teachers' instructional role while other anglophone cultures have drawn on constructivist and socio-cultural thinking to emphasise the role of the learner and informal classroom assessment.

Much has been claimed for the impact of AfL on learning. However the evidence base is restricted in terms of quantitative data, with most studies being small-scale and of a more qualitative nature. The evidence suggests an educationally significant impact, though of a more modest nature than has sometimes been claimed.

GENERAL CONCLUSIONS

Educational assessments affect what and how learning occurs. This is most obvious in classroom assessment, such as Assessment for Learning, where the theory and philosophy that motivates practice is rooted in the impact upon learning. In the other case that we discussed – international largescale assessments (ILSAs) – the connection between the assessments and learning is less direct because it is mediated through policy, curriculum and assessment design. ILSAs influence the zeitgeist, which as been termed 'soft governance', for its indirect but palpable effect upon policy (Bieber & Martens, 2011). We have argued that in both of these cases the alignment of assessment practice with a particular learning theory is less than straightforward. Thus, despite the impact of assessment upon learning, the use of learning theory to shape assessments is non-obvious.

Assessment design has often been motivated by technical considerations rather than a consideration of the likely impact of the regime upon learning. Relations between theories of learning and assessment design do not have a one to one mapping, but it would be true to say that cognitive theories are the current paradigm (Pellegrino et al., 2001). Practical and measurement considerations have come to the fore when the construct of interest is manifested in assessment artefacts. The philosophical assumptions behind educational assessments and their implications for practice are rarely recognised in the field and have resulted in stand-offs between camps. We have taken a postmodern pragmatic stance to educational assessment, arguing that constructs are socially constructed, that unidimensionality should not be prioritised over educational concerns and that invariance is a product of the nature of learning and the ways that we use assessment scores. In taking this position, we do not reject design quality concerns. Instead, we are arguing for the priority of educational concerns to a larger extent.

Use of educational assessment scores in the many second-order systems that have been generated for accountability purposes has created a gulf between what is assessed and how the data are used. What is signified by the assessments – specific forms of learning - and the signifier – for example international rankings – are disconnected. The signifier has taken on more significance than the signified in the case of ILSAs. Policy-makers who take decisions on the basis of educational assessment data rarely understand the content of the tests or the effects upon learning of changing them.

In this paper we have argued that educational assessment is a goal-setting activity and that it has a large impact upon the content and style of learning. Oftentimes, the negative effects of assessment upon learning have been noted in the literature, though we recognise that assessment can be motivational and that good assessment design can have positive effects. Designing tests worth teaching to has become a recognised goal (Popham, 1987; Resnick & Resnick, 1992; Linn, 2000; Shepard, 2000; Stobart, 2008). To do this, we argue that educational objectives would have to be prioritised over measurement where they are in conflict. Further, we argue that there needs to be greater focus upon the content and style of the assessments in terms of their likely washback. Use of assessment data only as signifiers is detrimental to this agenda.

The substantive learning theories discussed in this paper have been broad categories of theory. Of course, to truly motivate assessment design, learning theories would have to

be very specific regarding concepts, the order in which they should be taught in the curriculum and what should be considered as more valuable knowledge and skill. This remains a challenge because we are in the uncomfortable position of not being sure that the measurement can be extricated from the attribute of interest. We argue that learning theory must take into account the dynamic relationships between curriculum exposure, assessment design and learning outcomes.

Consider the volume of educational assessment that takes place internationally on an annual basis. It really is puzzling that we have not gained more from those data regarding students' learning. One reason is commercial. Testing is an industry that is risk-averse and proprietorial over data. Individual examiners will have gained craft knowledge over many years' experience with educational assessment data, but this has rarely been compiled and connected with learning theory. Equally, teachers will have built their own theories of action by using the information they gained from assessments, but again this has not been systematised. Thus, we argue that more open and systematic analysis of data is required to gain better understandings of the relationship between assessment and learning. This knowledge should be used to improve our learning and assessment theories, in the service of the design of tests worth teaching to. In other words, we should be seeking systemic validity, in which educational assessments bring about curricular, instructional and learning strategies that foster the cognitive traits that the assessments were designed to assess (Frederiksen & Collins, 1989).

REFERENCES

Alderson, J.C., and D. Wall. 1993. Does washback exist? Applied Linguistics 14, no. 2: 115–29.

Aldrich, J. (1995). Correlations genuine and spurious in Pearson and Yule. *Statistical Science*, *10*(4), 364 – 376. Retrieved from http://www.jstor.org/stable/2246135

Aldrich, J. (1995). Correlations genuine and spurious in Pearson and Yule. Statistical Science, 10(4), 364 – 376. Retrieved from http://www.jstor.org/stable/2246135

Ammermueller, A. (2007) Poor Background or Low Returns? Why Immigrant Students in Germany Perform so Poorly in the Programme for International Student Assessment, Education Economics, 15:2, 215-230, DOI: 10.1080/09645290701263161.

Anderson, J.R. (1990). Cognitive psychology and its implications. A series of books in psychology. WH Freeman/Times Books/Henry Holt & Co: New York. 3rd Edition. Xvi 519pp.

Anderson, J.R. (1996). ACT: A simple theory of complex cognition. *American Psychologist*, Vol 51(4), Apr, 355-365. http://dx.doi.org/10.1037/0003-066X.51.4.355.

Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms? *Medical Care*, *42*, I7–I16. doi:10.1097/01.mlr.0000103528.48582.7c

Andrich, D. (2014) A structure of Index and Causal variables. Rasch Measurement Transactions. 28, (3), 1475 - 1477.

Applebee, A. N., Langer, J. A. & Mullis, I.V.S. (1987) Learning to be Literate in America: Reading, Writing and Reasoning, Princeton NJ: Educational Testing Service.

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

Au, W. (2007). High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis. *Educational Researcher*, *36*, 258–267. http://doi.org/10.3102/0013189X07306523

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.

Baird, J. (2010). Beliefs and practices in teacher assessment. Editorial. Assessment in Education: Principles, Policy & Practice, 17, 1, 1 – 5.

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.A., Hopfenbeck, T.N., Newton, P.N., Stobart, G. & Steen-Utheim, A.T. (2014). Assessment and Learning. State of the Field Review. Oslo: Knowledge Centre for Education. Case Number 13/4697.

Bandura, A. (1969). Principles of behavior modification. New York: Holt, Rinehart & Winston.

Barsalou L.W. (1992) Cognitive Psychology: An overview for cognitive psychologists. Lawrence Erlbaum Associates, Hillsdale, NJ.

Bartlett, F.C. (1932, reprinted in 1977) Remembering: A Study in Experimental and Social Psychology. Cambridge: Cambridge University Press.

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

Berry, R. and Adamson, B. (2011) Editors. Assessemnt Reform in Education. Policy and Practice. Education in the Asia-Pacific Region: Issues, Concerns and Prospects. 14. Springer. Asia-Pacific Educational Research Association. UNEVOC International Centre.

Bieber, T. and Martens, K. (2011). The OECD PISA as a soft power in education? Lessons from Switzerland and the US. European Journal of Education Research, Development and Policy., 46, 1, 101 – 116. DOI: 10.1111/j.1465-3435.2010.01462.x

Biggs, J. (1998) Learning from the Confucian heritage: so size doesn't matter?
International Journal of Educational Research, 29 (8): 723 - 738.

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

Black, P. (2016) Formative Assessment implementation – and optimistic but incomplete vision. Assessment in Education: Principles, Policy and Practice, 22 (1) 161 – 177.

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, AprilBlinkhorn, S. F. (1997). Past imperfect, future conditional: Fifty years of test theory. *British Journal of Mathematical & Statistical Psychology*, *50*, 175–185.

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. (2003). 'In praise of educational research': formative assessment. *British Educational Research Journal*, 29, 5, 623 – 637. http://dx.doi.org/10.1080/0141192032000133721

Black, P., & Wiliam, D. (2006). Developing a theory of formative assessment. In J. Gardner (Ed.), *Assessment and learning* (pp. 81–100). London: Sage.

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

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

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

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

Borsboom, D. (2006a). The attack of the psychometricians. *Psychometrika*, *71*(451), 425–440. doi:10.1007/s11336-006-1447-6

Borsboom, D. (2008). Latent Variable Theory. *Measurement: Interdisciplinary Research & Perspective*, *6*(1-2), 25–53. doi:10.1080/15366360802035497

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

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.

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

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

Choppin, B. (1985). Lessons for Psychometrics from Themometry. *Evaluation in Education*, *9*, 9–12.

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.

Cronbach, L., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281 – 302.

Cronbach, L., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281 – 302.

Darling-Hammond, L. and Rustique-forrester, E. (2005), The Consequences of Student Testing for Teaching and Teacher Quality. Yearbook of the National Society for the Study of Education, 104: 289–319. doi:10.1111/j.1744-7984.2005.00034.x

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

Della Salla, S., & Anderson, M. (Eds.). (2012). Neuroscience in education: the good, the bad and the ugly. Oxford University Press.

Della Salla, S., & Anderson, M. (Eds.). (2012). *Neuroscience in education: the good, the bad and the ugly*. Oxford University Press.

Devitt, M. (1997). *Realism and truth.* Second edition. Princeton: Princeton University Press. 371 pages.

Dole, J. A.., Duffy, G.G., Roehler, L.R. and Pearson, P.D. (1991) Moving from the old to the new: Research on reading comprehension instruction. Review of Educational Research, 61 (2), 239 - 264.

Duggan, C. (2011). Dangerous and severe personality disorder. *The British Journal of Psychiatry*, May, 198, 6, 431 – 433.

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.

Eklöf, H (2010) Skill and will: test-taking motivation and assessment quality, *Assessment in Education: Principles, Policy & Practice*, 17 (4), 345–356.

El Masri, Y., Baird, J.A. and A. Graesser (2016) Language Effects in international testing: the case of PISA 2006 science items, *Assessment in Education: Principles, Policy & Practice*, Vol 23 (4) p XX.

Elley. W.B. (1992). How in the World do Students Read? IEA Study of Reading Literacy. International Association for the Evlauation of Educational Achievement. ISBN-92-9121-002-3. http://files.eric.ed.gov/fulltext/ED360613.pdf.

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. & Murphy, P. (2015). Editorial. Assessment in Education. Principles, Policy & Practice, 22(2), 1 – 5.

Engestrom, Y. (1987). Learning by expanding: an activity theoretical approach to developmental research. Helsinki, Orienta-Konsultit.

Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing*, 5(1), 23–35.

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.

Foucault, M. (1975). Discipline and Punish: the Birth of the Prison, New York: Random House. Available at: https://zulfahmed.files.wordpress.com/2013/12/disciplineandpunish.pdf. Accessed 31 July 2015.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. Educational Researcher, 18, 9, 27–32.

Freebody, P. & J.M. Freiberg (2011) The Teaching and Learning of Critical Literacy: Beyond the "Show of Wisdom", (p. 432 – 452) in Kamil, M.L., Pearson, P. D. , Moje, E.B. & P.P Afflerbach, (Eds) Handbook of Reading Research IV, Routledge, Taylor and Francis Group, New York.

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

Gagné, R.M. (1985). The Conditions of Learning and Theory of Instruction (4th Edition). New York, CBS College Publishing.

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).

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review*, *98*(2), 254–267. http://doi.org/10.1037/0033-295X.98.2.254

Gigerenzer, G., & Goldstein, D. G. (1996). GigerenzerGoldstein_MindAsComputer_CreativityResJ_1995.pdf. *Creativity Research Journal*, *9*(2&3), 131 – 144.

Gipps, C. (1994). Beyond Testing: Towards a Theory of Educational Assessment. London, Falmer

Goldstein, H. (1979). Consequences of Using the Rasch Model for Educational Assessment∗. British Educational Research Journal, 5(February), 211–220. doi:10.1080/0141192790050207

Goldstein, H. 2012. "Francis Galton, Measurement, Psychometrics and Social Progress." *Assessment in Education: Principles, Policy & Practice,* 19 (2): 147–158.

Goldstein, H. and Wood, R. 1989. Five Decades of Item Response Modelling. Journal of Mathematical and Statistical Psychology, 42: 139–16

Gould, S. J. 1996. The Mismeasure of Man. 2nd ed. New York, NY: W.W. Norton.

Graesser, A.C., K.K. Millis and R.A. Zwaan (1997), Discourse comprehension, Annual Review of Psychology, Vol 48, pp. 163 – 189.

Grisay, A (2003) Translation procedures in OECD/PISA 2000 international assessment, Language Testing, 20, 225–240

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

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

Guthrie, J. T. (2008) Engaging adolescents in reading, Corwin Press, Thousands Oaks, CA.

Guthrie, J. T. and A. Wigfield (2000) Engagement in Motivation in Reading, in M. L. Kamil & P. B. Mosenthal (eds), Handbook of reading research (Vol. 3, pp. 403 – 422), Erlbaum, Mahwah, NJ.

Guthrie, J. T., A. Wigfield, N.M. Humenick, K.C. Perencevich, A. Taboada and P. Barbarosa (2006) Influenes of stimulating tasks on reading motivation and comprehension.

Journal of Educational Research, 99, pp. 232 – 245.

Hacking, I. (2007). Kinds of People: Moving Targets. Proceedings of the British Academy, 151, 2006 Lectures. DOI:10.5871/bacad/9780197264249.003.0010.

Haladyna, T.M., & Rodriguez, M.C. (2013). *Developing and Validating Test Items*. Routledge, New York.

Hanson, F. A. (1994). Testing Testing: Social Consequences of the Examined Life. Berkeley, CA: University of California Press.

Harlen, W (2007) Criteria for evaluating systems for assessment, *Studies in Educational Evaluation*, 33, 15–28

Harlen, W. & M. James (1997) Assessment and Learning: differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice*, 4:3, 365 – 379.

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.

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. Journal of Applied Psychology, 92, 373-385.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? American Psychologist, May, 443–455.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? *American Psychologist*, *May*, 443–455.

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.

Hopfenbeck, T. N. and M. Kjærnsli (2016) Students' Test Motivation in PISA: The case of Norway. *The Curriculum Journal,* Vol. 27, NO. 3, 406 – 422.

Hood, S. B. (2008). Comments on Borsboom's Typology of Measurement Theoretic Variables and Michell's Assessment of Psychometrics as "Pathological Science." Measurement: Interdisciplinary Research and Perspectives, 6(1-2), 93–97. doi:10.1080/15366360802035554

Humphry, S.M. (2013). A middle path between abandoning measurement and measurement theory. *Theory & Psychology*, *23*, 770–785. doi:10.1177/0959354313499638

James, M. & Harlen, W.(1997) Assessment and Learning: differences and relationships between formative and summative assessment, Assessment in Education: Principles, Policy and Practice, 4, 3, 365-379 .

James, M. (2006) Assessment and learning. In: S Swaffield (ed.) *Unlocking Assessment. Understanding for reflection and application*. Abingdon, UK, Routledge, 20–35

James, M. (2012) Growing confidence in educational research: threats and opportunities, *British Educational Research Journal*, 38 (2), 181–201

James, W. (1890) *The Principles of Psychology*. New York, Henry Holt and Company.

Kane, M. (2008). The Benefits and Limitations of Formality. Measurement: Interdisciplinary Research and Perspectives, 6(1-2), 101–108. doi:10.1080/15366360802035562.

Kane, M. (2008). The Benefits and Limitations of Formality. *Measurement: Interdisciplinary Research and Perspectives*, *6*(1-2), 101–108. http://doi.org/10.1080/15366360802035562

Kellaghan, T., Madaus, G.F., & Airasian, P.W. (1982). The effects of standardized testing. Boston, MA: Kluwer-Nijhoff.

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.

Kuhn T. S. 1961. The function of measurement in modern physical science. Isis 52(168), 161-193.

Kingston, N. & Nash, B. (2011). Formative Assessment: A Meta-Analysis and a Call for Research, Educational Measurement: Issues and Practice, 30, 28–37.

Kingston, N. & Nash, B. (2015) Erratum *Educational Measurement: Issues and Practice* Summer 2015, Vol. 34 No. 2, p.55.

Kintsch, W. (1998). Comprehension: A paradigm for cognition. Cambridge, MA: Cambridge University Press.

Klenowski, V (2009) Assessment for Learning revisited: an Asia-Pacific perspective, Assessment in Education: Principles, Policy & Practice, 16 (3), 263–68

Klette, K., Lie, S., Ødegaard, M., Anmarkrud, Ø., Arnesen, N., Bergem, O.K., Roe, A. (2008) PISA +: Laerings- og undersvisningsstrategier i skolen {PISA+: Learning and teaching in Norwegian classrooms. Oslo: Norwegian Research Council.

Klette, K; Bergem, O. K. & A. Roe (2015). Teaching and Learning in Lower Secondary Schools in the Era of PISA and TIMSS. Springer Publishing Company. ISBN 978-3-319-17301-6. 195 s.

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.

Kreiner, S. and Christensen, K.B. (2014). Analyses of model fit and robustness. A new

look at the PISA scaling model underlying ranking of countries according to literacy. Psychometrika, 79, 2, 210 – 231. doi: 10.1007/s11336-013-9347-z

Laming, D. (1997). A critique of a measurement-theoretic critique: Commentary on Michell, Quantitative Science and the definition of measurement in psychology. *British Journal of Psychology*, 88, 389 – 391.

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.

Levine, M. (2000). Analytic Freud: Philosophy and Psychoanalysis. London: Routledge.

Lie, S., Kjaernsli, M. Roe, A., & A. Turmo (2001) Godt rustet for framtida? Norske 15 aaringers kompetanse i lesing og realfag i et internasjonalt perspektiv. Acta didactica, (4), Universitetet i Oslo.

Lewy, A (1996) Postmodernism in the field of achievement testing, *Studies in Educational Evaluation*, 22 (3), 223–244.

Linn, R.L. (2000). Assessments and accountability. Educational Researcher, 29, 2, 4 – 16.

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

Madaus, G.G. (1988). The distortion of teaching and testing: High-stakes testing and instruction. Peabody Journal of Education, 65(3), 29-46. DOI: 10.1080/01619568809538611

Madaus, G F, Russell, M K, & 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 Inc.

Maraun, M. D. (1998). Measurement as Normative Practice. Theory & Psychology, 8(4), 435–461.

Maraun, M. D. (1998). Measurement as Normative Practice. *Theory & Psychology*, *8*(4), 435–461.

Mari, L., Maul, A., Irribarra, D. T., & Wilson, M. (2013). Quantification is Neither Necessary Nor Sufficient for Measurement. *Journal of Physics: Conference Series*, *459*, 012007. doi:10.1088/1742-6596/459/1/012007.

Markus, K A & Borsboom, D (2013) *Frontiers of Test Validity Theory: Measurement, causation and meaning.* New York, Routledge.

Maul, A. (2013a). Method effects and the meaning of measurement. Frontiers in Psychology, 4(April), 169. doi:10.3389/fpsyg.2013.00169.

Maul, A. (2013a). Method effects and the meaning of measurement. *Frontiers in Psychology*, *4*(April), 169. http://doi.org/10.3389/fpsyg.2013.00169

Maul, A. (2013b). On the ontology of psychological attributes. Theory & Psychology, 23, 752–769. doi:10.1177/0959354313506273

Maul, A. (2013b). On the ontology of psychological attributes. *Theory & Psychology*, *23*, 752–769. http://doi.org/10.1177/0959354313506273

McNamara, T. (2001). Language assessment as social practice: challenges for research. Language Testing, 18(4), 333–349. doi:10.1191/026553201682430076

McNamara, T. (2001). Language assessment as social practice: challenges for research. *Language Testing*, *18*(4), 333–349. http://doi.org/10.1191/026553201682430076

Mesic, V. (2012) Identifying Country-Specific Cultures of Physics Education: A differential item functioning approach, International Journal of Science Education, 34 (16), 2483–2500.

Messerli, F. H. (2012). Chocolate consumption, cognitive function, and Nobel Laureates. The New England Journal of Medicine, (367), 1562 – 1564. Retrieved from DOI: 10.1056/NEJMon1211064

Messerli, F. H. (2012). Chocolate consumption, cognitive function, and Nobel Laureates. *The New England Journal of Medicine*, (367), 1562 – 1564. Retrieved from DOI: 10.1056/NEJMon1211064

Michell, J. (1997a). Quantitative science and the definition of measurement in psychology. British Journal of Psychology, 88, 355–383. doi:10.1111/j.2044-8295.1997.tb02641.x.

Michell, J. (1997a). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*, 355–383. http://doi.org/10.1111/j.2044-8295.1997.tb02641.x.

Michell, J. (1997b). Reply to Kline, Laming, Lovie, Luce and Morgan. British Journal of Psychology, 88(1997), 401–406. doi:10.1111/j.2044-8295.1997.tb02647.x

Michell, J. (1997b). Reply to Kline, Laming, Lovie, Luce and Morgan. *British Journal of Psychology*, *88*(1997), 401–406. http://doi.org/10.1111/j.2044-8295.1997.tb02647.x

Michell, J. (2005). The logic of measurement: A realist overview. Measurement: Journal of the International Measurement Confederation, 38(2005), 285–294. doi:10.1016/j.measurement.2005.09.004

Michell, J. (2005). The logic of measurement: A realist overview. *Measurement: Journal of the International Measurement Confederation*, *38*(2005), 285–294. http://doi.org/10.1016/j.measurement.2005.09.004

Michell, J. (2008a). Is Psychometrics Pathological Science? *Measurement: Interdisciplinary Research and Perspectives*, *6*(1-2), 7–24.

http://doi.org/10.1080/15366360802035489

Michell, J. (2008b). Rejoinder. *Measurement: Interdisciplinary Research and Perspectives*, 6(1-2), 125–133. doi:10.1080/15366360802121917

Michell, J. (2008b). Rejoinder. *Measurement: Interdisciplinary Research and Perspectives*, *6*(1-2), 125–133. http://doi.org/10.1080/15366360802121917

Michell, J. (2008c). Rejoinder. *Measurement: Interdisciplinary Research and Perspectives*, *6*(1-2), 125–133. doi:10.1080/15366360802121917

Michell, J. (2013). Constructs, inferences, and mental measurement. *New Ideas in Psychology*, 31(1), 13–21. doi:10.1016/j.newideapsych.2011.02.004

Michell, J. (2013). Constructs, inferences, and mental measurement. *New Ideas in Psychology*, *31*(1), 13–21. http://doi.org/10.1016/j.newideapsych.2011.02.004

Mislevy, R. J. (1997). Postmodern Test Theory. In A. Lesgold, M. J. Feuer, & M. B. Allison (Eds.), *Transitions in Work and Learning: Implications for Assessment* (pp. 180 – 198). Washington, D.C.: National Academy Press.

Mislevy, R. J. (1997). Postmodern Test Theory. In A. Lesgold, M. J. Feuer, & M. B. Allison (Eds.), *Transitions in Work and Learning: Implications for Assessment* (pp. 180 – 198). Washington, D.C.: National Academy Press.

Mislevy, R.J., Steinberg, L.S. & Almond, R.G. (2003). Focus Article: On the Structure of Educational Assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 1, 3-62.

Moss, P.A., Pullin, D. Gee, J.P. & Haertel, E.H. (2005). The idea of testing: psychometric and sociocultural perspectives. *Measurement: Interdisciplinary research and perspectives*, 3, 2, 63 – 68.

Moss, P. A., Pullin, D. C., Gee, J. P., & Haertel, E. H. (2007). Assessment , Equity , and Opportunity to Learn. Learning in Doing: Social, Cognitive and Computational Perspectives. Edited by.

Nardi, E. (2008) Cultural biases: a non-Anglophone perspective, Assessment in Education: Principles, Policy & Practice, 15 (3), 259–266.

Neisser, U. (1967). Cognitive Psychology. New York: Appleton Century Crofts.

Newton, P.E. 2007. Clarifying the purposes of educational assessment. Assessment in Education: Principles, Policy & Practice, 14(2): 149–170.

Newton, P. E., & Shaw, S. E. (2014). *Validity in Educational and Psychological Assessment*. London: Sage Publications Ltd.

Oancea, A (2005) Criticisms of educational research: key topics and levels of analysis, British Educational Research Journal, 31 (2), 157–183

OECD (2000) Measuring Student Knowledge and Skills The PISA 2000 Assessment of Reading, Mathematical and Scientific Literacy Education and Skills. OECD, Paris.

OECD (2001). Knowledge and Skills for Life. First results from PISA. Paris: OECD Publishing.

OECD (2009). PISA 2009 Assessment Framework: Key Competencies in Reading, Mathematics and Science. Paris: OECD Publishing.

OECD (2010a) Viewing the United Kingdom School System through the Prism of PISA (www.oecd.org/pisa/46624007.pdf).

OECD (2010b) PISA 2009 Assessment framework – Key competencies in reading, mathematics and science. OECD, Paris.

OECD (2014). PISA 2012 Technical Report. http://www.oecd.org/pisa/pisaproducts/pisa2012technicalreport.htm. Accessed 22 September 2015.

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.

Olsen, R.V., Prenzel, M. & Martin, R. (2011). Interest in science: a many-faceted picture painted by data from the OECD PISA study. Editorial. Special Issue: Students' Interest in Science across the World: Findings from the PISA study. *International Journal of Science Education,* 33, 1 – 6.

Pearson, K. (1982). *The grammar of science*. London: Dent.

Pellegrino, J. W., Chudowsky, N., and Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

Popham, W J (1987) The merits of measurement-driven instruction, Phi Delta Kappan, 68 (9), 679–82.

Popham, W.J. (2008) Transformative Assessment. Association for Supervision and Curriculum Development (ASCD), Alexandria, VA

Porter, M. (1986). The rise of statistical thinking, 1820 - 1900. Princeton, NJ: Princeton University Press.

Porter, M. (1986). *The rise of statistical thinking, 1820 - 1900.* Princeton, NJ: Princeton University Press.

Prais, S. J. (2004). Cautions on OECD's recent educational survey (PISA), Oxford Review of Education, 29 (2), 139–163.

Pressley, M., Goodchild, F., Fleet, J., Zajchowski, R., & Evans, E.D. (1989) The Challenges of classroom instruction, Elementary School Journal, 89, (3) , 301 - 342.

Pressley, M. (2006) What the future of reading research could be. Paper presented at the International Reading Association Reading Research Conference, Chicago, IL.

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.

Pryor, J. & Crossouard, B. (2008). A socio-cultural theorisation of formative assessment, Oxford Review of Education, 34 (1), 1–20.

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

Roos, B. & D. Hamilton (2005) Formative assessment: A cybernetic viewpoint. *Assessment in Education: Principles, Policy and Practice*. 12:1, 7 -20.

Rumelhart, D. E. (1998). A connectionist approach. In Mind Readings: Introductory Selections on Cognitive Science. Massachusetts: Massachusetts Institute of Technology.

Rumelhart, D. E. (1998). A connectionist approach. In *Mind Readings: Introductory Selections on Cognitive Science*. Massachusetts: Massachusetts Institute of Technology.

Rutkowski, L. & Rutkowski, D. (2016). A call for a more measured approach to reporting and interpreting PISA results. *Educational Researcher*. ISSN 0013-189X. *45*(4), s 252-257 . doi: 10.3102/0013189X16649961.

Sadler, R.D. (2007) Perils in the meticulous specification of goals and assessment criteria, *Assessment in Education: Principles, Policy & Practice*: 14:3, 387 – 392.

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.

Schleicher, A. (2009) International Benchmarking as a Lever for Policy Reform. In Hargreaves, A. & M. Fullan (Eds) Change Wars. Bloomington, IN Solution Tree.

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 PhilosophicalAssociation*, 64 (3), 21–37 (www.jstor.org/stable/3130074).

Shepard, L A (2000) The role of assessment in a learning culture, Educational Researcher, 29, 7, 4–14.

Simon, H. A. (1979). *Models of Thought*. (Yale University Press, Ed.). London.

Simon, H. A. (1996) Observations on the Sciences of Science Learning. Paper prestned for the Committee on Developments in the Science of Learning for the Sciences of Science Learning: An Interdisciplinary Discussion. Department of Psychology, Carnegie Mellon University.

Sijtsma, K. (2006). Psychometrics in Psychological Research: Role Model or Partner in Science? *Psychometrika*, *71*(3), 451–455.

Skinner, B.F. (1953). Science and human behavior. New York: Macmillan.

Solano-Flores G, Backhoff E & Contreras-Nino, L A (2009) Theory of test translation error, International Journal of Testing, 9 (2), 78–91 Searle, J.R. (1990). Is the brain's mind a computer program? Scientific American, January, 26 – 31.

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.

Stenner, J., Burdick, D.S., & Stone M.H. (2008). Formative and reflective models: can a Rasch Analysis tell the difference? Rasch Measurement Transactions, 22(1), 1152-53.

Stenner, J., Stone, M.H., & Burdick D.S. (2009). Indexing vs. measuring. Rasch Measurement Transactions, 22(4),1176-77.

Sternberg, Robert J. 1981 Testing and cognitive psychology. American Psychologist, Vol 36(10), Oct, 1181-1189. http://dx.doi.org/10.1037/0003-066X.36.10.1181

Stiggins, R J (2005) Student-involved assessment for learning. Upper Saddle River, NJ, Prentice Hall.

Stobart, G. (2008). Testing Times. Routledge, Abingdon, UK.

Stobart, G. (2012) Validity in formative assessment. In J.Gardner (Ed.) *Assessment and Learning,* London: Sage).

Stobart G. & Eggen, T. (2012) High-stakes testing – value, fairness and consequences, Assessment in Education: Principles, Policy & Practice, 19:1, 1-6, DOI: 10.1080/0969594X.2012.639191

Stobart, G., Elwood, J., & Quinlan, M. (1992). Gender Bias in Examinations: how equal are the opportunities? British Educational Research Journal, 18(3), 261–276. doi:10.1080/0141192920180304

Stobart, G., Elwood, J., & Quinlan, M. (1992). Gender Bias in Examinations: how equal are the opportunities? *British Educational Research Journal*, *18*(3), 261–276. http://doi.org/10.1080/0141192920180304

Swaffield, S. (2011) Getting to the heart of authentic Assessment for Learning, *Assessment in Education: Principles, Policy & Practice*, Vol 18, Issue 4, 433 – 449.

Taras, M (2009) Summative assessment: the missing link for formative assessment, *Journal of Further and Higher Education*, 33 (1), 57–69.

Tesio, L. (2014a) Causing and Being Caused: Items in a Questionnaire May Play a Different Role, Depending on the Complexity of the Variable. Rasch Measurement Transactions, 28(1), 1454-56.

Tesio, L. (2014b) Items and Variables, Thinner and Thicker Variables: Gradients, not Dichotomies. Rasch Measurement Transactions. 28, (3), 1477 – 1479.

Thurstone, L.L. (1959). *The Measurement of Values.* Chicago: University of Chicago Press.

Torrance H. (2012). Formative assessment at the crossroads: conformative, deformative and transformative assessment. Oxford Review of Education, 38 (3), 323-342.

Torrance, H. (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 & Pryor, J (1998) *Investigating Formative Assessment. Teaching, Learning and Assessment in the Classroom*. Buckingham, Open University Press.

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, J.B. (1930). Behaviorism. New York: Norton.

Webb, J. (Ed.). (2013). Voyages of discovery. In *Nothing. From absolute zero to cosmic oblivion - amazing insights into nothingness*. London: Profile Books Ltd.

Wechsler, D. (1944). Measurement of adult intelligence, (3rd edm). Baltimore, MD: Williams and Wilkins.

Wieman, C E (2014) The similarities between research in education and in the hard sciences, *Educational Researcher*, 43 (1), 12–14

Wiliam, D. (2008) International comparisons and sensitivity to instruction, Assessment in Education: Principles, Policy a&nd Practice, 15 (3), 253 – 257.

Wiliam, D. (2010) *An integrative summary of the research literature and implications for a new theory of formative assessment*, In Andrade, H.L. & G. J. Cizek (Eds) Handbook of Formative Assessment, Routledge, New York.

Wiliam, D (2011) What is assessment for learning?, Studies in Educational Evaluation, 37 (1), 2–14.

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 AssessmentSystem, *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.

Wolming, S., & Wikstrom, C. (2010). The concept of validity in theory and practie. Assessment in Education. Principles, Policy & Practice, 17(2), 117 – 132.

Wolming, S., & Wikstrom, C. (2010). The concept of validity in theory and practie. *Assessment in Education. Principles, Policy & Practice*, *17*(2), 117 – 132.

Wu, M (2005) The role of plausible values in large-scale surveys, *Studies in Educational Evaluation*, 31, 114–128.

Wu, A.D. and Ercikan, K. (2006). Using multiple-variable matching to identify cultural sources of differential item functioning. International Journal of Testing, 6, 3 287 – 300.

Yin, R.K. (2009). Case Study Research. Design and Methods. Fourth Edition. Applied Social Science Research Methods Series. Volume 5. California: Sage.

Young, M.F.D. (2008). Brining Knowledge Back In. From social constructivism to social realism in the sociology of education. Routledge: Abingdon, UK.

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