Journal of Education and Practice
ISSN 2222-1735 (Paper) ISSN 2222-288X (Online)
Vol.4, No.12, 2013

What Makes a Good Secondary Assessment? On Achieving the Aims of Assessment

Daniel P. Hyde
Doctorate of Education student, School of Social Science & Public Policy, King's College London, Strand, London WC2R 2LS, UK
* E-mail of the corresponding author: daniel.hyde@kcl.ac.uk

Abstract
Drawing on the wealth of testing literature, and considering recent assessment paradigm shifts, this paper outlines four main purposes of assessment: as a measure of achievement and attainment; as a gate-keeping selection tool; to promote a meritocracy, providing equality of opportunity, and; to keep schools and teachers accountable. Through a critical discussion of traditional and alternative assessments, tensions between equality, accountability, assessment and the curriculum are highlighted in an attempt to find what would make a 'good' assessment – a test that achieves all four aims. Ultimately, it is suggested that the only way to achieve this is to reconceptualise assessment and curriculum as two halves of the same whole. Formative and summative assessment shows students' competency. A curriculum that supports deep-learning, monitored to be gender-neutral, supports gate-keeping and equality. Finally, making schools accountable through qualitative inspections of their teaching and curriculum removes the negative-effects of teaching-to-the-test.
Keywords: assessment, authentic assessment, accountability, curriculum, equality

1. Introduction

The literature on secondary assessment is vast, covering many facets and inviting a wealth of debate. We see articles and papers on what assessments should achieve. Assessment models and methods are evaluated in light of these aims. Different problems are discussed in relation to these methods - for example gender and assessment, how assessments affect learning, accountability and assessment etc. This paper addresses these problems together, highlighting tensions between the different aims of assessment, in an attempt to discover what would make a 'good' assessment – one which achieves all the stated aims.

Assessment has four main purposes. The first is to find out what students know and can do. Whilst doing this, it can judge the curriculum and provide information to improve teaching practice and student achievement through self-evaluation. By establishing what it is a student is capable of, it performs a gate-keeping function to allow or deny access to higher education or professions. Thirdly, it promotes equality by providing a level-playing field for students of any background or gender to show their abilities and achievements. Finally, in recent years it has become a way of making schools and teachers accountable to governments, education authorities and parents.

This paper addresses different methods of assessment by placing them in two categories. One category covers traditional assessments which comprise mainly of multiple-choice, standardised tests popular in the United States, and the summative written papers sat in the UK. The second covers alternatives which, alongside written papers, include work assessed formatively in a classroom context, and demonstratively through practical displays of skills. I will show that the traditional methods, particularly short-answer and multiple-choice questions seem only to work as a cost-efficient accountability measure. Alternatives will be shown to work much better as a measure of knowledge and capability and therefore as gate-keepers.

However, by looking at all the purposes of assessment together, we will see that there are problems with equality and accountability. It is difficult to obtain an equal and subjective treatment of students with alternative assessments, and when measures are put into place to achieve this goal, tensions arise with students' learning and knowledge. It will be argued that criteria-reference marking, put in place to introduce objectivity, can lead to a narrowing of students' subject knowledge and a focus on achieving a test goal rather than obtaining a thorough understanding of the subject.

Accountability has a similarly negative effect on teaching and learning, as the pressure of league-table comparisons 'pushes teachers into 'teaching to the test' rather than 'teaching for understanding'' (Harrison, 2007, p. 215).

It would seem that no assessment method can manage to completely fulfill all the purposes we see assessment to have, as the aims begin to work against each other. This paper suggests that it is necessary to see the curriculum as a distinct part of the process, having its own role to play in achieving these aims. If assessments focus on the display of students' knowledge and capabilities, and their competency for future work, the curriculum could prom ote equality through teaching practice. Finally, this paper will propose that accountability is centred on the process rather than the product of education. By making schools and teachers accountable purely through the inspection process, rather that the quantitative measure of examination results, accountability could lead to

improvements in practice rather than the detrimental effects on student learning this paper will outline. When examination league-tables were introduced following the 1988 Education Act, parents' access to data on schools was limited. Twenty-five years on, 80% of households have access to the internet (ONS, 2012). The possibilities for parents to access qualitative, detail rich information on schools through the online catalogue of school inspection reports make inspection a more effective way of keeping schools accountable.
A good assessment is not impossible. We may have to change the way we view our criteria - and understand that the purposes of assessment can be achieved through the curriculum as well as the testing methods themselves.

2. The purpose of assessment

Before we can enter into a discussion evaluating assessment methods, it is necessary to outline the various aims and purposes of assessment in general. Across associated literature (Harrison, 2007; Eisner, 1993; Royce Sadler 1994; DeLuca *et al.*, 2012), there are numerous stated and wide-ranging reasons for teachers, schools and societies to assess students. Here, we shall outline them as four main purposes.

Purpose 1: Achievement and attainment

First and foremost, the purpose of assessment is to measure student achievement and attainment. In doing this, it accomplishes a number of goals. It finds out what students have learnt, 'determining what learners know or what they can do' (Lum, 2012, p.590). Even if we acknowledge the ontological problem of whether it is ever possible to identify truly what a person knows, we can use assessment to indicate student behaviours of competence (*ibid*, p595). It establishes a measure of whether a student is fit for the workforce, whether in terms of vocational skills or, indeed, general literacy and numeracy – 'important competencies... needed in today's society' (Palm, 2008, p.3). This in turn, informs whether course outcomes have been met, particularly for courses which are intended to provide specific skills, abilities and techniques.

By providing a judgement on the success of learning objectives, assessment allows feedback on the success of teaching and curricula. Assessment is a useful tool in teacher self-reflection, allowing them to improve practice. As Eisner (1993) states, 'good teaching and substantive curricula cannot be mandated; they have to be grown' (p. 224). Eisner, Deluca et al. (2012) and Palm (2008) each document the capability of assessment results to inform and support that process.
In simple terms, the above can be described as 'assessment of learning'. Similarly, self-reflection through assessment is possible for students. The Assessment Reform Group in the UK have spearheaded the use of 'assessment *for* learning', which uses formative testing to allow students to gain a greater insight into the way they are assessed with the aim of improving their attainment.

Purpose 2: Gate-Keeping

The second main purpose of assessment is for gate-keeping. By establishing students' knowledge and competencies, and by allowing comparison between candidates, assessment provides a gate-keeping function for entry into further or higher education, professions or societies. The gate-keeping function has played a pivotal role in shaping assessments and syllabi, not only vocationally for entering professions such as law, accountancy and medicine, but at school level. The British 'gold-standard' school A-Level certificate's primary focus on its inception was to facilitate entry to undergraduate courses – highlighted by the prominent role played by university examination boards in writing syllabi and setting examinations. The shift in power over the curriculum from universities to examination boards has meant the usefulness of A-level as a university selection tool (and, therefore, as a gate-keeper) has come into doubt. Universities, one of the key 'gate-keeping' stakeholders of A-Levels, perceive 'students lack a number of... skills they require at first-year level. Many students appear not to have a wide contextual understanding' (Hyde, 2010, p. 50). It seems that the importance of the gate-keeping function will be central to A-Level reforms in the years to come (Ofqual, 2012).

Purpose 3: For equality

Another purpose of assessment is to promote equality. In simple terms, tests can be argued to provide an equal experience for students across all social strata as each student is measured using the same criteria. Assessment can be (and, as the discussion on methodology and equality will show, this cannot be a given) a way of fairly comparing students and their achievements. If assessment could be seen to be an accurate measurement of students' merit, then there are key equality issues here. Merit-based assessment 'stemmed from the desire to eliminate access on the grounds of social class or status, patronage, wealth or blatant corruption' (Royce Sadler, 1994, p.116). An assessment could be said to be fair if all students are treated the same and the test allows all students to achieve. The *outcome* does not have to be equitable across students. It would negate the purpose of gate-keeping or comparison if every student achieved the same standards. Marrying the equality of opportunity with a conception of equity which 'demands fair competition but tolerates and, indeed, can require unequal results' (Espinoza, 2007) leads us to a measure of fairness which creates little tension with gate-keeping.

Purpose 4: Accountability

Finally, beyond student- and teacher-centred aims, assessment provides society, both nationally and internationally with a general 'educational temperature-taking' (Eisner, 1993, p.223). Standardised Attainment Tests (SATs) in both the US and the UK provide information on the numeracy and literacy skills of citizens, not just to give results on specific students, schools, local education authorities etc., but to inform education policy decisions at a governmental level.

This general 'temperature-taking' sowed the seed for a pervasive regime of accountability. As Eisner (1993) documents, what started as a nationwide American concern regarding a perceived decline in standards suggested by a fall in SATs scores in the early 1970s led to state-mandated assessment programmes where 'testing became a focal point of educational reform' (Harnisch & Mabry, 1993, p.180). Whilst the Curriculum Reform Movement (CRM) in the US argued for a change in assessment practice that would support teacher development, promote a wide variety of activities for students and place at its core the quality of the content of the curriculum (Eisner, 1993), the foci of government agencies were: ease of measurement; economic data capture, and; comparability across students, schools and states.

These contested foci of education policy could also be seen in the UK. At the same time as the CRM argued for a move away from the standardised tests favoured by government, the New Sociology of Education in the UK strove for a focus on the qualitative. They put at the centre of their work societal issues compounded by education policies and practices: in essence defining the problems with the system rather than providing neat measures on which the government could base policy (Shain & Ozga, 2001). All the while the state was reacting to problems in the system by increasing accountability and measurement. This marketisation of education (sparked by the 1988 Education Act) included the introduction of the National Curriculum and league tables so parents could compare schools. On the one hand, the National Curriculum could be seen as an equality-based policy ensuring all students received a standardised curriculum. Alternatively, it could be seen as state-controlled mandating of the curriculum and assessment, setting minimum attainment levels allowing schools and local education authorities to be easily (yet rather primitively) compared. Despite fears that 'tests were driving the curriculum in unwanted directions' (Harnisch & Mabry, 1993, p.180), the subsequent two decades have seen no move away from the need to measure student attainment for the purpose of comparison and target-setting: compounded with the New Labour fixation on the 'virtue of accountability' (Barber, 2004).

3. Methods of Assessment

To attempt to define what might make a good assessment, the discussion must also concern the variety of methods we use in assessing students and how effectively they meet the purposes stated above. Different assessment methods fulfil (and hinder) these purposes and there are very definite tensions at play between the different aims of assessment.

3.1 Traditional Assessments

So-called 'traditional' assessments have been the mainstay of the US and UK education systems for decades. Whilst contrary voices can be heard as far back as the 1960s (Schwab, 1969), the use of multiple choice 'Standardised Attainment' tests and examination essay questions (in the US and the UK respectively) has survived through a whole host of criticism in the literature (Bagnato & Macy, 2010; Cumming & Maxwell, 1999; Linn, Baker, & Dunbar, 1991; Palm, 2008). There are a number of reasons for the survival of traditional tests, mainly concerned with issues of equality and accountability.

One undoubted strength of traditional testing is that it is a straightforward and cost-effective way of collecting data. This makes it popular with governments and other agencies charged with monitoring school performance. One of the important features of accountability in education is so that the public 'can see the benefits of their investment' (Barber, 2004, p.12). A non-multiple choice test can cost two-hundred times more to mark than a multiple-choice paper (Wiggins 1990). It seems axiomatic that a system part-designed to audit value for money should be cost-effective. Also, the straightforward results of multiple choice questions inherent in a right/wrong mark scheme mean that comparison and statistical analysis is simple across an entire nation.

However, whilst there are some positive elements of standardised tests, they do not support a number of the stated purposes of assessment outlined above. If we look at our first purpose of assessment: the need to measure student achievement and attainment, it is true that traditional tests can claim to demonstratively cover a far wider amount of the curriculum. A multiple choice test may contain hundreds of diverse questions across a syllabus whereas alternative methods to traditional tests (discussed below) might only be able to cover very specific sections (Harnisch & Mabry, 1993; Macintosh & Morrison, 1969). If we are to test students' capabilities and knowledge, the question must be asked whether these kinds of tests really tell us what a student knows.

Whilst at a philosophical level, as already noted, it can be argued that we can never truly know what another person actually knows, there are degrees to which we can observe their behaviour to make a judgement regarding their competency. Lum (2012) outlines two concepts of assessment: prescriptive and expansive. Whilst it is too simple a reading of his work to define traditional tests as 'prescriptive' and alternative assessments as

'expansive', one-word answers and multiple-choice questions are undoubtedly prescriptive. They look for one 'right' answer rather than any understanding, or not, of what might be behind the answer. The problem here is in what Lum defines as the 'right/wrong' scenario. It is entirely possible for a candidate to give the 'right' answer to the question yet have a total lack of understanding of the topic being tested. We could ask more questions to check further understanding, but ultimately, '*however many questions are set* it is still logically possible that the Right/Wrong scenario could obtain' (Lum, 2012, p.596): there could always be a vital misunderstanding which the test misses. Therefore, traditional multiple-choice and one word or short answers with specific mark-schemes do not suit the purpose of assessing a students' capability or understanding. This ultimately would suggest that these assessments are incapable of working effectively as a gate-keeping function, as it seems likely that the profession or university etc. would be looking to accept candidates proficient in their vocation or subject, not ones merely able to answer prescriptive questions 'correctly'.

Just as the questions do not give an infallible judgement of knowledge, the results of multiple choice papers and short-answer questions do not give useful feedback that could support improved performance and self-reflection amongst teachers and students (Eisner, 1993; Wiggins 1990). Harnish & Mabry (1993) go further to suggest that standardised tests not only fail in this regard, they succeed in 'limiting and misdirecting instruction' (p. 186).

For a similar reason, traditional tests fail our third purpose: the promotion of equality. Perhaps the most cited and compelling case for the US SATs is that they are by their very nature standardised. Multiple choice questions provide one correct answer. They are therefore seen to be a reliable and fair measure of a student's capability. There can be no possibility of subjectivity by the marker – indeed, in the US the marker is an optical-mark recognition (OMR) machine obviously incapable of being anything other than objective. Written one-word answers marked by an examiner using a proscriptive mark-scheme provide the same benefit. For a fair comparison between students, tests must be 'carried out in the same way for all individuals, scored in the same way, and the scores interpreted in the same way' (Gipps, 1994, p.284). Whilst acknowledging that there are degrees of objectivity, and there are still subjective forces at play with such tests (for example, the choice of subject content and how questions are worded), Macintosh and Morrison (1969) believe 'it follows, therefore, that subjectivity in the marking of scoring of the test is eliminated' (p. 9).

However, subjectivity in the marking process disguises the real equality issue here. As discussed above, because the method does not allow the student to show the process behind their answer, it is 'quite possible for a student to arrive at a correct answer for entirely irrelevant reasons' (Eisner, 1993, p.227). It hardly seems fair for students who have a solid, deep understanding of a subject to be outscored by someone who has guessed the correct answer – and this is a possibility with some traditional tests. Whilst it is claimed the standardised nature of the tests supports concepts of equality and fairness, students may ironically feel somewhat unfairly treated when being compared using such a fallible measure of knowledge.

Traditional tests fail to meet the needs of three of our aims of assessment. They do not accurately measure student achievement and attainment, which therefore makes them unfit gate-keepers. They also do not promote equality for candidates. Being the only purpose this method achieves, it seems likely its accountability benefits underpin its popularity with policy makers.

3.2 Alternative Assessment

The deficiencies of the format, and therefore results, of traditional tests (devised to easily compare students' achievements) led to the development of alternative assessment models. Across the literature, there are descriptions and definitions of 'performance assessment' and 'authentic assessment'. Whilst there are differences between 'performance' and 'authentic' assessment, they suit discussion together here as they ultimately prioritise similar goals. To discuss them, I shall use the generic 'alternative assessment'. By avoiding the term 'authentic assessment', we escape the semantic argument regarding the validity of the word. In one of the few criticisms of authentic assessment, Terwilliger (1997) finds fault with the method because its name is 'misleading and confusing. The term inappropriately implies that some assessment approaches are superior to others because they are more "genuine" or "real"' (*ibid.* p. 27). For a genuine discussion of the benefits and disadvantages of a method it is important to go beyond linguistics.

The common priorities of alternative assessments are neatly described by Palm (2008), falling into three categories he defines as: life beyond school; curriculum and classroom practice, and; learning and instruction.

1. Life beyond school – assessments look to skills and cognition required by adults and aim to mirror them. This includes specific work related tasks but also conditions that mirror adult work-life situations, such as time constraints.
2. Curriculum and classroom practice – this is a broader view of authenticity which expects assessment to mirror what happens in the classroom. Whilst it may seem 'inauthentic' in the sense that it doesn't necessarily mirror real-life, its authenticity lies in assessing performance in a context familiar to the child's learning environment and is an antidote to the multiple choice test or the timed-essay summative exam. It also demonstrates the spectrum of authenticity which can

be seen in alternative assessments.

3. Learning and instruction – assessment may be regarded as authentic if it provides formative feedback and allows Assessment for Learning (Black & Wiliam, 1998).

3.3 Alternative assessments and students' knowledge and capabilities

When judging alternative assessments against our four purposes outlined earlier, they certainly seem to fare well on the aim of testing students' knowledge and capabilities. This is because one benefit of alternative assessments is that they test competence in skills in a demonstrative rather than an abstract way. The tasks are 'real examples of the skill or learning goals rather than proxies as in standardised tests' (Gipps, 2010, p. 284) emulating 'kinds of mastery demonstrated by successful adults'(Cumming & Maxwell, 1999, p.178). They are more likely to result in 'deep' rather than 'surface' knowledge – a concept outlined by Marton & Säljö (1976) which identifies the understanding of a concept and the ability to apply it across a variety of circumstances against a mere repetition of a learnt idea.

Alternative assessments are more likely to result in deep-level processing for a number of reasons. Firstly, in preparation for assessment 'the focus is more likely to be on thinking to produce an answer than on eliminating wrong answers as in multiple choice tests' (Gipps, 2004, pp. 284-285). Secondly (and this is more effective when the assessment is conducted formatively in the classroom context), the teacher does not need to leave the teaching of a concept to engage in an 'indirect or proxy' distraction (Wiggins, 1990). The 'proxy' in this case being the exam, something which teachers need to prepare for in its own right, above and beyond the time spent on teaching the concept. Thirdly, the formative element of the assessment promotes self-reflection and evaluation of understanding promoting an engagement with the subject beyond a mere surface level.

Because the assessments mirror real-life tasks, they are a more reliable measure of competency. Using as an example an engineer designing a bridge, Cumming & Maxwell (1999) state:

> For performance of such complexity there is little comparison with traditional expectations of school instruction and assessment, where the emphasis has been on the production of predefined 'correct' answers to predefined closed 'problems'. Similar considerations apply to all kinds of expertise, for example bricklaying, acting, tennis, hairdressing, musical performance and welding. (pp. 181-182).

If one of the purposes of education is to prepare children for the world of work and a purpose of assessment, as discussed above, is to measure whether they are fit for the workforce, it is important that tests assess competency and encourage mastery of skills. This call has been echoed by those in skills-based work, most recently by the UK employers' organisation, the Confederation of British Industry (CBI, 2012).

Recognising that alternative assessments work within the field of a child's learning experience, one may make the point that what is an authentic situation for one student may not be authentic for another. One may also question what 'real-life' situation would be appropriate for non-vocational subjects such as history (Terwilliger, 1997). However these challenges miss the point that these assessments are an alternative to traditional testing – and cannot answer the question of how a multiple-choice or timed essay can ever be 'real'.

The answer to this problem could lie in an understanding of there being degrees of authenticity. Notwithstanding Terwilliger's problem with the definition of 'authenticity', across the literature 'authentic' and 'performance' assessment are often used interchangeably, or occasionally with a simple differentiation of where the assessment takes place - in a classroom setting or an exam room. As stated earlier, it can be argued that there is a spectrum of assessment, from standardised through to entirely performance-based, and across these there are degrees of authenticity. For example, in simple terms, the question "What is 1+3?" is a purely standardised test based on a mathematical concept. If we change the question to "If we add 1 apple to a bag containing 3 apples, how many apples will we have in total?", does this make the question more 'authentic'? It is certainly more context-based. Whilst we would not define the question as 'authentic' or 'alternative' in terms of assessment, it does illustrate that there are degrees to be found across assessments. Different subjects have different requirements and display authenticity in different ways. Terwilliger questioned how history could be assessed 'authentically'. The answer lies in the degrees of authenticity used. It makes sense that at least 60% of GCSE and A-Level Drama exams use an examined practical performance as an assessment (Edexcel, 2010). This brings an authenticity to the subject and puts the competencies learnt into a vocational context.

As Palm's 'Learning and instruction' classification might suggest, alternative assessment demonstratively provides for assessment for learning. It facilitates students' self-reflection and evaluation, allowing them to improve their work and hone their abilities. It does this by the use of clear criteria referencing. To assess work beyond a simple 'is this right or wrong?' method (inherent in multiple-choice assessment) requires specific criteria for assessors which can be shared with teachers and students. To enhance student attainment, 'assessment processes and procedures need to be clear and transparent to students' (thu Vu & Dall'Alba, 2010, p. 1). If students understand the criteria with which they are to be assessed, they are able to facilitate their own learning and improvement. This is at the heart of Black & Wiliam's (1998) work on assessment for learning which has since been accepted as good practice in UK schools (DCSF, 2008).

3.4 Alternative assessments as a gate-keeper

Thus as alternative assessments seem to be a good measure of achievement and attainment, it should stand that these methods also fulfil a gate-keeping purpose. There are reasons why these methods work well to this aim. Alternative assessments are more than just essay-based tests. They can include coursework, portfolios and oral presentations. These can provide the mirror of a 'real-life' situation, as:

> In many colleges and all professional settings the essential challenges are known in advance – the upcoming report, recital, Board presentation, legal case, book to write, etc. Traditional tests, by requiring complete secrecy for their validity, make it difficult for teachers and students to rehearse and gain the confidence that comes from knowing their performance obligations. (Wiggins, 1990)

They also allow a student to display the best possible work they can produce which could be argued is a more accurate measure of their capabilities than a traditional exam. Gipps (1994) makes the link with Vygotsky's *zone of proximal development* (ZPD). Coursework assessments where the teacher plays a supportive role assisting and guiding the student through the work may result in the best performance rather than a typical performance of a child's abilities but it also reflects the scaffolding required for a child's cognitive development in Vygotsky's theory. If assessment is to be used as a predictor of future ability (as it undoubtedly is when used as a gate-keeper) then it seems perfectly acceptable for assessments to be based on 'the upper limits of their ZPD' (Hohenstein & King, 2007, p.178). Whilst alternative assessments with criterion-referenced marking usually steer clear of the peer-comparison which is at the heart of norm-referenced marking in traditional assessments, Wiggins (1990) makes an interesting observation that if all the students know the criteria on which they are being marked, and what is expected of them, then it is possible to hold them to higher standards as they all should be working at a higher level. The same can be said of teacher assistance. Whilst it could be suggested that assessing students based on 'scaffolded' work potentially lowers standards, Wiggins' argument could be used in this example – so long as we are expecting the best work possible of our candidates, then if we compare them we can increase our expectations as a result.

Following such a list of positive facets, it would seem obvious that a good assessment would have to follow the model of the alternative assessments described above. However, they are not without criticism.

3.5 Alternative assessment and equality

Our third purpose of assessment is its promotion of equality. If students are to be compared it is essential that the process is fair and equitable. Alternative assessments as outlined above show a large amount of variety due to a lack of standardisation. This can be manifested in a number of ways. Firstly, it does not seem impossible that the administration of the same task can vary from centre to centre. However, the problem with non-standardised tests is that the reliability of such tests has been shown to be questionable as there is evidence of 'great variation in administration not only across teachers but also between administrations by the same teacher' (Gipps, 2010).

The criteria-reference marking discussed above as a benefit to promote assessment for learning has been used as an answer to the standardisation problem. Specific criteria are given to examiners to judge performance (whether practical, oral or written) and these same criteria are used in examining all students. Not only does this still not solve the problem that different examiners may interpret criteria in different ways but it has also been suggested that criteria-reference marking may actually encourage teacher-examiners to be over-confident in their marking and move beyond the criteria and be influenced by internalised decisions on the candidates (Hay & Macdonald, 2008). This could lead to unfair enhancement or degradation of students' achievements.

Criteria-reference marking also causes tensions with the ability of alternative assessments to succeed in their main purpose of measuring student achievement. The move from holistic mark-schemes to specific mark-schemes which list what should be observed to gain credit in an assessment has led to a 'mechanistic approach to subjects' (Tomlinson, 2004, p. 87). This then promotes a narrower subject knowledge as students are taught to the test, knowing they will gain credit for making the expected points rather than being free to display their knowledge. This results in "'thin skills" rather than "rich knowledge"' (Davis, 1995, quoted in Lum, 2012, p.590). Lum's 'expansive' method of assessment, in which assessors use all the observable evidence at their disposal to make a holistic judgement of whether a student is competent, is also denigrated by specific marking criteria. They turn the assessment into a prescriptive method which allows for the 'right/wrong' scenario which is a problem with traditional tests.

The problem of the product (assessment) negatively influencing the process (teaching, learning and the curriculum) - as seen with the effect of criteria-reference reference marking - can also be seen when looking at the use of demonstrations in the classroom mirroring work within a vocational context. Whilst multiple-choice or short answer questions may affect the curriculum and teacher practice, encouraging drilling of correct answers without independent creative thought and a deep understanding, alternative assessments may also distort teaching and learning.

Whilst the argument has been made for assessing work in a more authentic context – in a classroom whilst students are working, through observed performances, or whilst undertaking the skill for which a competence is

to be judged – it could lead to a greater use of situated learning. Situated learning occurs when concepts are learnt through a direct, contextual example (see Anderson *et al.* 2000; (Cumming & Maxwell, 1999). It could be argued that situated learning is the 'process' equivalent to the above alternative assessment 'product'.

It is possible that not only does situated learning lead to alternative assessments such as portfolios, observations, performance-based tests (or, in other terms, a kind of situated assessment), but the reaction could happen the other way round. Some alternative assessments could lead to purely situated learning – a similar effect to multiple-choice questions leading to learning by drills. In this case, students are taught purely by practical example to accommodate the assessment.

This may seem a positive. After all, we are aiming to assess competency. But we also have established the need for a deep understanding of concepts. Indeed, it was the drive of the CRM in 1960s America (in response to the Soviet lead in the space race) to want 'students to understand the structure of mathematics, not just be competent at computation' (Eisner, 1993, p.221). One of the criticisms of situated learning is that it limits transfer of learning. Anderson *et al.* (1996) use the example of Brazilian street sellers being adept at fast arithmetic in their vocational setting, but unable to perform abstract calculations in a school context. Cumming & Maxwell (1999) cite research showing the differential between adults' mathematical performing in a supermarket setting compared with a classroom environment.

This argument is more against a situated-only focused pedagogy rather than alternative assessments. Just as Anderson, Greene and Reeder *et al.* (2000) found in their study with children aiming darts at underwater targets, knowledge transfer was more effective when concepts were taught first, and then contextual practice was given. So long as the curriculum and teaching styles allow for conceptual learning, subsequent context-based assessments may actually support the understanding process.

It is important that curriculum and pedagogy distinguish between the process and the product, or 'learning' goals and 'performance' goals (Dweck 1989 cited in James & Pedder, 2006). James & Pedder compare the two by differentiating learning goals, where students aim to gain new knowledge or a new skill, with performance goals which are about behaviour purely to gain 'favourable judgements of... their competence' (James & Pedder, 2006). Terwilliger (1997) is also concerned that this assessment merely leads students to gain observable performative skills rather than the cognitive understanding of the concepts behind them. What these critiques miss is that they highlight the distinction of learning and assessment as two different things. So long as it is possible to ensure teaching and curricula focuses on understanding and process rather than assessment and product, there does not need to be a problem. Indeed, one may question whether assessment can ever measure 'learning' goals. If assessment is an observation of their competence (through whichever method), it is always going to be a measure of their performance.

Ultimately, when looking at their ability to assess student achievement and attainment, alternative assessments work better than traditional tests overall. However, it does raise the issue that when striving for one purpose, here equality, a method may influence the provision of the curriculum – the product might negatively influence the process.

3.6 Alternative assessment and accountability

Looking at the final purpose of assessment we find a more prosaic criticism of alternative assessment - their cost. To fulfil assessment's purpose of accountability, as a measure of national achievement both generally and comparatively between students, tests must be carried out under the auspices of one main agency (eg. Government, whether delegated to examination boards etc.). Therefore, standardisation processes need to be put into place to ensure an appearance of equality (even if the above criticisms can still be made about whether they are truly equal or standardised). This is costly. Even in a paper promoting the use of alternative assessments, Wiggins (1990) admits that 'the scoring of judgement based tasks seems expensive when compared to multiple-choice tests (about $2 per student vs. 1 cent)'.

4. Achieving the 'good' assessment

If we are to arrive at what would be a 'good' assessment, by which we mean one that fulfils all the purposes of assessment outlined above, we need to look more closely at what is successful and unsuccessful with these methods.

As traditional tests only perform well in terms of accountability, and accountability itself has a mandate to 'improve outcomes for all students' (Barber, 2004), it seems that good assessment cannot be found by using traditional tests alone. However, alternatives do not promote equality because of a lack of standardisation and their questionable 'fairness'. Neither method, for the reasons stated above, can be shown to truly fulfil an equality aim. This leads us to go back to our purposes, and look more closely at our requirement for equality. Whilst this paper is not going to argue that students should not expect equality of opportunity in education, perhaps it is where and how this equality is achieved that could hold the answer. There are many equality issues in education, social economic status, race, sexual orientation etc., certainly too many to address here. But many of the general

equality issues can be discussed by using one equality consideration. Here we shall look at gender in assessment. It has long been understood that some methods of assessment can seem to be biased towards males or females. Multiple-choice questions have been seen to favour boys, whereas it has been demonstrated that girls outperform boys in coursework (see Chilisa, 2000; Cole, 1997). It would seem sensible to aim for an assessment which balanced different methods. Even if this was possible, and the assessment was still fit for the other purposes outlined, Chilisa outlines further gender bias, not just in the method but in the writing of questions and choice of subject matter. Her study, looking at examinations in Botswana, highlights questions which might favour boys or girls – for example, questions about the male reproductive system or the female menstrual cycle. This argument fails to consider that in a comprehensive biology curriculum we may wish students to know about male and female biology. Testing one or the other, rather than both, can be argued to keep the test more unknown, promoting more thorough learning. Students would need to understand the whole syllabus rather than just aim to achieve on questions that favour their gender. We might need to ensure that a paper didn't solely cover one sex, but this is more of an issue with examination content generally. I have already stated earlier the need for tests to represent an assessment of the whole curriculum, not just small sections.

Chilisa also outlines a problem with examination texts that use pronouns 'he', 'him' and 'his' more than female equivalents. A specific examination example highlights a literature comprehension text which contains the sexist treatment of women. If we were to excise all literature from our English or History examination curricula that did not fit with our twenty-first century sensibilities, we may end up with a very narrow education, and many classic texts condemned. However, Chilisa's point is very useful here because both issues raised relate to the curriculum. We might find the study of 'The Taming of the Shrew' misogynistic if it was taught with no commentary, or discussion of the context in which the play was written and how attitudes may now be different. If we begin to look at the curriculum and the way it is delivered as a key, but potentially separate, part of assessment, then our measure of assessment becomes a little easier. Stobart et al. (1992) claims 'examinations shape the curriculum and so which aspects of a subject are assessed and how this is done will play a part in making an examination fair.' Perhaps, instead, we could think that whilst examinations *do* shape a curriculum, it is therefore essential that we look at how it is then *delivered and taught*, to make assessment overall more fair. For example, if it is the case, as the studies suggest, that girls outperform boys in coursework assessments, we can address this by ensuring more pedagogical support for boys during the coursework process. The equality issue is resolved by a change in curriculum delivery rather than abandoning the assessment model altogether.

This separation of process (curriculum and teaching) and product (assessment) could also be used with our other sticking-point: accountability. In his compelling case for accountability in education, Michael Barber makes the point that 'many of the arguments about accountability systems are inevitably and rightly about *how* rather than *whether* [schools should be accountable]' (Barber, 2004) . Furthermore, he questions 'if' tests are to be involved (rather than making it an axiomatic assumption that they are. If tests are not involved we are left with inspection – a measure of process rather than product. If accountability was a product of a nation's desire to ensure value for money, it might make sense to leave accountability to the existing inspection system (albeit perhaps redesigned) which saves the money on number-crunching simple comparisons. With wider access to the internet allowing parents access to every inspection report on all schools in the country this still facilitates the marketisation aspect which league-tables sought to achieve.

The other negative impact of assessment on the curriculum was the effect of criteria-reference marking on deep knowledge. Again, this is a problem which can be addressed by acknowledging the curriculum and its delivery as separate from the assessment. If accountability measures are to be focused on curriculum delivery then what happens in schools becomes even more important. The inspection system allows for lesson observations and scrutiny of pupils' work. If criteria for the demonstration of deep-level understanding in the process was expected it may discourage the negative impact of assessment.

5. Conclusion

Synthesizing points raised across the wide amount of relevant literature, this paper outlines four main purposes of assessment: to find out what students know and can do; whether they meet competences for further work and/or study; to promote a meritocracy and ensure equality of opportunity, and; to ensure schools are accountable to parents and governments. Traditional assessments give a vast amount of quantitative data at very little cost, and therefore they have been a focus of the accountability culture of the last thirty years. However, they have very few positive attributes when compared with alternatives. Alternative assessments, formative and summative, including oral tests, coursework, practical demonstrations, and written papers which allow for a range of answers and a display of a breadth and depth of knowledge, allow students to demonstrate what they know and can achieve. This allows them to be a better indicator of their capabilities beyond school. Problems occur when considering equality and accountability. The tensions between gender, accountability, assessment and the curriculum suggest a more nuanced approach should be taken, looking at the potential of achieving the

purposes of assessment through both the curriculum and assessments.

Assessments cannot be divorced from the curriculum, and there is no suggestion here that we see them detached. Seeing them as being separate parts of the same whole means that all four purposes of assessment can be achieved through the alternative methods discussed above. Firstly, students' knowledge and competences can be shown through a variety of formative and summative tests assessed through a range of means: portfolio, oral, written tests in a classroom and an examination context. Secondly, a gate-keeping purpose is achieved by ensuring that the curriculum promotes deep-learning and skills required by interested stakeholders (for example the CBI and universities). Thirdly, the curriculum should promote equality. State agencies (such as Ofqual which regulates and accredits qualifications in the UK) can ensure varied syllabi which are gender neutral. Teacher training and observed lessons can ensure teachers address equality issues in the classroom process. Finally, schools and teachers can still be accountable qualitatively through the detailed inspections which occur regularly in every school in the country. These inspection reports, now widely accessible by the majority of parents, provide information on the success or otherwise of a school's process – and give rich feedback for teachers to improve their practice. If accountability is totally embedded in a qualitative judgement of the curriculum and its teaching, rather than weighted towards quantitative assessment, the four aims of assessment can be achieved.

References

Anderson, J. R., Reder, L. M., Simon, H. A., Herbert, L. M. R., & Simon, A. (1996). Situated Learning and Education. *Educational Researcher*, 25(4), 5–11.

Anderson, J. R., Greeno, J. G., Reder, L. M., & Simon, H. A. (2000). Perspectives on Learning, Thinking and Activity, *Educational Researcher,* 29(4), 11–13.

Bagnato, S. J., & Macy, M. (2010). Authentic Assessment in Action: A "R-E-A-L" Solution. *NHSA Dialog: A Research-to-Practice Journal for the Early Childhood Field*, 13(1), 42–45.

Barber, M. (2004). The Virtue of Accountability : System Redesign , Inspection, and Incentives in the Era of Informed Professionalism, *Journal of Education*, 185(1), 7–38.

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

Brainard, M. B. (1997) Assessment as a way of seeing, in Goodwin, A. L. (1997) ed. *Assessment for Equity and Inclusion*, Routeledge, London.

Bridgeman, B. & Schmitt, A. (1997). Fairness issues in test development and administration, in Willingham, W. & Cole, N. (1997). *Gender and fair assessment,* Lawrence Erlbaum Associates, New Jersey.

CBI. (2012). *First Steps: A new approach for our schools*, CBI, London

Chilisa, B. (2000). Towards Equity in Assessment: Crafting gender-fair assessment, *Assessment in Education: Principles, Policy and Practice*, 7(1), 61-81

Cole, N. (1997), Understanding gender difference and fair assessment in context, in Willingham, W. & Cole, N. (1997). *Gender and fair assessment*, Lawrence Erlbaum Associates, New Jersey.

Cumming, J. J., & Maxwell, G. S. (1999). Contextualising Authentic Assessment. *Assessment in Education: Principles, Policy & Practice,* 6(2), 177–193.

DCSF, (2008). *The Assessment for Learning Strategy*, DCSF, accessed on 02 January 2013 from https://www.education.gov.uk/publications

DeLuca, C., Chavez, T., & Cao, C. (2012). Establishing a foundation for valid teacher judgement on student learning: the role of pre-service assessment education. *Assessment in Education: Principles, Policy & Practice*, 1–20.

Dillon, J. & Maguire M., eds. (2007) *Becoming a Teacher: Issues in Secondary Teaching*, Open University Press, Maidenhead.

Edexcel. (2010). *Edexcel Advanced Subsidiary GCE in Drama and Theatre Studies Specification.* Edexcel Limited.

Eisner, E. W. (1993). Reshaping assessment in education : some criteria in search of practice. *Journal of Curriculum Studies*, 25(3), 219–233.

Elwood, J. (1995). Undermining Gender Stereotypes: examination and coursework performance in the UK at 16, *Assessment in Educational Principles, Policy & Practice*, 2(3), 283-303

Elwood, J. (1999). Gender, achievement and the 'Gold Standard': differential performance in the GCE A level examination. *Curriculum Journal,* 10(2), 189-208.

Espinoza, O. (2007). Solving the equity–equality conceptual dilemma: a new model for analysis of the educational process. *Educational Research*, 49(4), 343–363.

Gipps, C. (1994). Developments in Educational Assessment : what makes a good test ?, *Assessment in Education: Principles, Policy & Practice,* 1(3), 283–291.

Goodwin, A. L. (1997) ed. *Assessment for Equity and Inclusion*, Routeledge, London.
Harrison, C. (2007) Making assessment work in the classroom, in Dillon, J. & Maguire M., eds. (2007*) Becoming a Teacher: Issues in Secondary Teaching,* Open University Press, Maidenhead.
Hay, P. J., & Macdonald, D. (2008). (Mis)appropriations of criteria and standards-referenced assessment in a performance based subject. *Assessment in Education: Principles, Policy & Practice,* 15(2), 153–168.
Heywood, J. (2006) The Formal Assessment of Student Learning: Alternative Assessment, in *Engineering Education: Research and Development in Curriculum and Instruction*, John Wiley & Sons, Inc., Hoboken, NJ, USA. doi: 10.1002/0471744697.ch16
Hohenstein & King (2007). Learning in and outside of the classroom, in Dillon, J. & Maguire M., eds. (2007*) Becoming a Teacher: Issues in Secondary Teaching*, Open University Press, Maidenhead.
Hyde, D. (2010) *Curriculum 2000 Ten Years On: A case study on perceptions of preparedness of A-Level music students for university study*, dissertation submitted for the MA in Education Management, King's College London.
Jablonka, E. V. A., & Gellert, U. W. E. (2012). Potentials, Pitfalls, and Discriminations: Curriculum Conceptions Revisited, *New Directions in Mathematics and Science Education*, 23, 287–307.
James, M., & Pedder, D. (2006*).* Beyond method: assessment and learning practices and values. *Curriculum Journal,* 17(2), 109–138.
Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, Performance-Based Assessment: Expectations and Validation Criteria*. Educational Researcher,* 20(8), 15–21.
Lum, G. (2012). Two Concepts of Assessment*. Journal of Philosophy of Education*, 46(4), 589–602.
Macintosh, H. & Morrison, B. (1969). *Objective Testing*, University of London Press, London.
Marton F. and Säljö R. (1976). On qualitative differences in learning. *British Journal of Educational Psychology,* 46: 4–11.
Ofqual (2012). Ofqual announces changes to A-Levels, Retrieved on 13 January 2013 from http://www.ofqual.gov.uk/news/ofqual-announces-changes-to-a-levels/
ONS. (2012). Statistical Bulletin Internet Access - Households and Individuals , 2012 (pp. 1–19). Retrieved 13 January 2013 from http://www.ons.gov.uk/ons/dcp171778_275775.pdf
Palm, T. (2008). Performance Assessment and Authentic Assessment : A Conceptual Analysis of the Literature. *Practical Assessment, Research & Evaluation,* 13(4).
Royce Sadler, D. (1994). Examinations and Merit. *Assessment in Education*, 1(1), 115–120.
Rust, C., Price, M & O'Donovan, B. (2003). Improving students' learning by developing their understanding of assessment criteria and processes, *Assessment and Evaluation in Higher Education,* 28: 2, 147 – 164.
Schwab, J. (1969*). College Curriculum and Student Protest*, University of Chicago Press, Chicago.
Shain, F., & Ozga, J. (2001). Identity Crisis ? Problems and Issues in the Sociology of Education Identity Crisis ? Problems and Issues in the Sociology of Education, *British Journal of Sociology of Education*, 22(1), 109–120.
Slaughter, S. (1997*).* Class, race and gender and the construction of postsecondary curricula in the United States: Social movement, professionalization and political economic theories of curricular change, *Journal of Curriculum Studies*, 29:1, 1-30.
Stobart, G., Elwood, J., & Quinlan, M. (1992). Gender Bias in Examinations : how equal are the opportunities ? *British Educational Research Journal*, 18(3), 261–276.
Terwilliger, J. (1997). Research news and Comment: Semantics, Psychometrics, and Assessment Reform: A Close Look at "Authentic" Assessments. *Educational Researcher*, 26(8), 24–27.
thu Vu, T. & Dall'Alba, G. (2010). Identifying authentic assessment practices for student learning, *Australian Association for Research in Education*, Retrieved 15 November 2012 from: http://ocs.sfu.ca/aare/index.php/AARE/AARE_2010/paper/view/2105
Tomlinson, M. (2004) *14-19 Curriculum and Qualifications Reform: final report of the working group on 14-19 reform*, London, DfES
Wiggins, G. (1990*).* The case for authentic assessment. *Practical Assessment, Research & Evaluation, 2(2).* Retrieved 12 November 2012 from http://PAREonline.net/getvn.asp?v=2&n=2
Willingham, W. & Cole, N. (1997). *Gender and fair assessment*, Lawrence Erlbaum Associates, New Jersey.

This academic article was published by The International Institute for Science, Technology and Education (IISTE). The IISTE is a pioneer in the Open Access Publishing service based in the U.S. and Europe. The aim of the institute is Accelerating Global Knowledge Sharing.

More information about the publisher can be found in the IISTE's homepage: http://www.iiste.org

CALL FOR PAPERS

The IISTE is currently hosting more than 30 peer-reviewed academic journals and collaborating with academic institutions around the world. There's no deadline for submission. **Prospective authors of IISTE journals can find the submission instruction on the following page:** http://www.iiste.org/Journals/

The IISTE editorial team promises to the review and publish all the qualified submissions in a **fast** manner. All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Printed version of the journals is also available upon request of readers and authors.

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar