CAMBRIDGE ASSESSMENT

A consideration of assessment validity in relation to classroom practice

Nat Johnson
Cambridge Assessment

A paper presented at the International Association for Educational Assessment Annual Conference, Baku, Azerbaijan, September 2007.

Nat Johnson
Evaluation Group
Research Division
Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU
United Kingdom
Direct dial. +44 (0)1223 553839
Fax. +44 (0)1223 552700
Email: johnson.n@cambridgeassessment.org.uk

www.cambridgeassessment.org.uk

Cambridge Assessment is the brand name of the University of Cambridge Local Examinations Syndicate, a department of the University of Cambridge.

Cambridge Assessment is a not-for-profit organisation.

Abstract

High stakes assessment has the potential to promote misdirected effort by both test takers and their teachers. Teachers may spend a significant amount of time working through past test material to prepare their students for assessments.

Test focussed teaching can result in both meaningful gains in test scores, indicating a greater grasp of the subject content, and undesirable gains resulting from coaching in common question styles without an increase in content understanding. Any increase in test scores achieved through coaching without a sound understanding of the subject content presents a threat to the validity of the assessment.

If it is accepted that ‘teaching to the test’ is inevitable in a high stakes setting then it is important that we investigate the extent to which different item styles are susceptible to superficial coaching strategies as well as to consider the test context.

This paper looks at the validity of questions in relation to how they are used in the classroom for test preparation. The paper explores the validity theory and classroom practice pertinent to test focussed teaching and reports on the development of a multi-item scale which aims to identify the extent to which a given question encourages the unproductive aspects of ‘teaching to the test’.

Introduction

In exploring the ideas surrounding validity and classroom practice the range of topics covered includes many of the current concerns in assessment, from ‘teaching to the test’ and ‘coaching’, to implications affecting pedagogical thinking and policy implications regarding the burden that assessment places on schools.

In his opening keynote address to the last IAEA conference (McGaw, 2006) Barry McGaw commented upon the power of assessment and its potential in a high stakes setting for promoting misdirected effort. His observations that “system level assessment, far from driving positive reform, can lead to mis-directed effort” and that high-stakes assessment “results in a focus in teaching on the outcomes that can be measured rather than those that are important” are representative of a growing unease regarding the effect of assessments on educational practice, particularly in the classroom.

In the UK there has been a great deal of public debate surrounding the issues of the accomplishment of students moving on from A-level to degree level or employment. The UK press has for some years engaged in highly critical reporting on an annual basis at the time of the publication of GCSE and A-level results (public assessments taken at age 16 and 18 respectively). This coverage frequently cites employers and those responsible for university admissions complaining that students have narrow abilities and cannot apply their knowledge outside of the assessment situation (Warmington & Murphy, 2004).

The problem of low general competency perceived by employers and universities whilst many high grades are awarded is not inherently contradictory. Those closely involved with candidates' scripts suggest that in some cases the standard of what pupils produce under examination conditions is getting higher and that the higher grades are deserved. Chief Examiners report that "markers noted that overall candidate performance was slightly better than previous entries" (OCR English language A level, Chief Examiner's Report, 2007a) and that "the general standard of the candidates was a little higher than in previous years" (OCR Mathematics A level, Chief Examiner's Report, 2007b). Critics suggest that it may then be the case that the pupils have a narrow competency range which only covers the exam material and not what they (the critics) consider to be important within the subject. Additionally, candidates display different profiles of skills in attaining a grade and a candidate's particular strengths may not be as a stakeholder desired, leading to further criticism. It is difficult to establish empirical provenance for claims in this emotive area, as many arguments are simply over what content is valued by different stakeholder groups. However (Gipps, 1994) suggests that in some situations teachers are under pressure to change their approach, commenting that "It is not that teachers want to narrow their teaching, nor to limit unduly students' educational experience, but if the test scores have significant effects on people's lives, then teachers see it as part of their professional duty to make sure that their pupils have the best possible chance they can to pass the test".

This is the perceived problem of 'coaching' or 'teaching to the test'. Candidates who have been 'taught to the test' may gain the qualification but not possess the knowledge that some may expect from an individual gaining the qualification. Indeed, it is so endemic that Gipps (1994) comments "Teaching to the test is a relatively well understood activity in the UK, although here it might be called preparation for examinations". This problem has possibly been accentuated by the perception that national accountability pressures on schools develop a culture where schools need good results to maintain their intake and funding. A recent report referring to the UK National Curriculum tests suggests that "league tables turn the tests into high stakes assessment. The unfortunate side effects of this can include teaching a narrow and shallow form of the curriculum tailored to the test" (The Advisory Committee on Mathematics Education, 2002). The specific issue that undermines an assessment in the eyes of stakeholders is that if a student has simply learnt to answer the question through practice (or through specific coaching techniques geared to successfully answering that question, rather than general teaching of skills and content), and has not mastered the underlying principles then the contribution of the assessment to society and the economy is undermined.

However, a word of caution is required, as the effect of assessment on the curriculum is not, as some may suggest, wholly negative, with the effect of certain accountability measures delivering gains in performance that promote improvements in teaching leading to valid gains in attainment. The study of the comparability of national test standards in the UK national tests between 1996 and 2001 suggested that the implementation of the testing regime had been successful in levering up standards of attainment in England (Massey, Green, Dexter & Hamnett, 2003). The evaluation also suggested that the positive benefits of the accountability measures imposed had been widespread, finding that "since the advent of national tests, achievement levels in schools have in fact improved substantially in almost all curriculum areas/key stages investigated".

Exploration of literature on validity

The concerns outlined refer to misdirected effort in high stakes assessment and how it can lead to stakeholder confidence in an assessment being undermined. The challenge created by perceptions that candidates are unable to do the things stakeholders believe that they are certified to do is a challenge to the validity of the assessment

Formal definitions of validity continually develop. The ideas considered here begin with (Messick, 1993) and move through various critics (Mehrens, 1997) to new approaches to validity (Kane, 2006) which are based in legal argument to build a case for the validity of the assessment. Much of the early validity work relates to objective testing, which is not frequently used within the UK tradition. Additional ideas surrounding the validity of performance assessment in extended writing (Moss, 1992) and authenticity of task (Torrance, 1995) are of interest and possibly more suited to an examination based system, and Messick (1996) recognises them as "tacit validity standards".

A useful starting point when considering modern validity is Messick's opening statement in his seminal chapter on validity in the third edition of 'Educational Measurement'.

"Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores." (Messick, 1993)

This central definition summarises some thirty years' work by Messick and others as they attempted to define the boundaries of validity. However, it became clear that some elements of the quality or usefulness of assessments did not lie within the remit of the formal validity that had been devised. These were elements of the use of an assessment that lay outside the examination room, in the classroom and the wider society, and therefore outside the direct control of the assessment developer. This need for a wider remit for evaluation led to the development of other validity concepts, namely consequential, curricular and systemic validities.

- *Consequential validity* refers to what happens to assessment results when used and interpreted in society. It is summarised by Shephard (1997) as "the incorporation of test consequences into validity investigations". The question of to what extent consequential validity is an independent concept is a subject of much discussion. Consequential validity has never truly stood alone as an isolated conceptualisation of validity, nor was it intended to be so. Messick (1993) describes "the social consequences of testing as an integral part of validity", suggesting that it is not a stand-alone validity concept, but an important component of validity evaluation.

- *Curricular validity* as originally introduced by McClung (1978, cited in Wood, 1991) refers to the correspondence between what the examination assesses and the objectives of instruction, However, more recently it has begun to be used interchangeably with consequential validity, prompting discussion about which ideas should be included within validity study, to which we return later.

- Frederiksen and Collins (Frederiksen & Collins, 1989) developed the idea of *Systemic validity*, defining a systemically valid test as

  > *"one that induces in the education system curricular and instructional changes that foster the development of the cognitive skills that the test is designed to measure".*

  With its cognitive focus on what is occurring within the examination room Frederiksen and Collins (Frederiksen & Collins 1989) concept of systemic validity proves useful. An assessment exhibits poor systemic validity by inducing different classroom behaviours from those that the syllabus writers intended. If a candidate answers a question purely through familiarity with the question style, rather than through an understanding of the

> subject content then it could be argued that they have not developed the cognitive skills that the test is designed to measure. This concept has remained influential, with the expression of cognitive skills being particularly helpful in considering the concept.

In an assessment with high systemic validity, an experienced practitioner in the subject would be expected to do well, because what is being tested is the subject, and not how the individual copes with formalised (contrived) scenarios for assessment. Music examinations could be considered to have high systemic validity. They are also an extreme example of teaching to the test, in that the candidate prepares a number of pieces of music with their tutor, with interpretation given by the tutor. There is no suggestion that this style of assessment does not result in individuals displaying (and largely retaining) the assessed skills.

To be useful, the concepts within systemic validity need to be focussed in a narrow area. Systemic validity does not cover the breadth of social interpretations of results; visible and measurable aspects such as the right people getting jobs and teaching programmes having an impact on community development, which might be considered to fall within consequential validity. Systemic validity include less measurable ideas; curricular alignment, pedagogical change and teaching to the test.

These three ideas of consequential, curricular and systemic validities, which relate to the wider effects of assessment have been presented as ‘validities’, but questions have been raised as to whether or not they are part of formal validity evaluation.

Shephard (1997) argues that the proliferation of ‘validities’ is an unnecessary confusion. In relation to consequential validity she reasons that “By coining a new term, antagonists and advocates have created a false impression that a new kind of validity was invented....it is true that our understandings of validity theory have evolved... [they] are not outside the underlying network of relationships that frame a validity investigation.” Shephard is at pains to note that consequential considerations are part of construct validity, but involve value judgements. Moss (1992) discusses “overburdening the concept of validity to the point where it ceases to provide useful guidance”.

Messick (1995) reduced the proliferation of validities by incorporating the wider validity concepts within a new six faceted construct validity which included the conceptual elements of most previous contributions. He clarified the standing of consequential validity, making it one of the facets, retaining it within formal validity investigations, but not as a stand alone concept. In other writings, Messick (1996) makes the point that "it is not that adverse social consequences of test use render the use invalid but, rather, that adverse social consequences should not be attributable to any source of test invalidity such as construct–irrelevant variance." Construct irrevelant variance is the variation in candidates' marks that does not come from their knowledge of the ideas being tested. It can result in candidates gaining fewer marks than their knowledge would suggest (if some part of the question presentation disadvantages them). Equally it can lead to the award of more marks then their knowledge alone would deserve (for example if they recognise a question type or have additional help through cheating).

(Mehrens, 1997) argued that the remit of validity investigations was becoming too large, arguing that "The issue is not whether to analyze effects of a particular application but whether to call that a validity investigation." Mehrens suggests "reserving the term for determining the accuracy of inferences about the characteristic being assessed, not the efficacy of actions following assessment". In this paper it seems appropriate to consider consequential validity and curricular validity as sources of threats to validity rather than as central parts of the validation process.

This is a view strengthened by more recent work. Validation work has recently moved towards 'validity argument' (Kane 2006). This approach developed within legal assessment and is a process whereby evidence for the validity of the assessment is provided qualitatively, but in a robust and substantiated way. This more holistic view is a counterpoint to the more quantitatively driven validation practice of the objective test tradition in the USA and is perhaps more suited to the examination-based assessment tradition in the UK. The ability of assessment agencies to identify threats to validity and to use that information to change their practice then becomes an important part of a new formal validity.

Practical approaches to the study of the effects of assessments have been developed within other fields of study. Two terms from the language testing community provide helpful additional concepts.

- *Washback* is a term that is widely used within the language testing community and is defined by Hamp-Lyons (1997, cited in Hawkey, 2006) to "refer to a set of beliefs about the relationship between testing and teaching and learning". Washback occurs at a micro level affecting the actions of individual learners and teachers.

- Another term from language testing is *Impact*, which is an approach to study and evaluates "the net effects of a programme" (Weiss, 1998). Impact studies focus on longer-term development (particularly relevant when considering the collateral benefits of language education in developing countries). Impact is considered at a macro level involving wider stakeholders and systems.

The ways of thinking about validity explored above provide a framework within which actions outside of the control of test developers are considered as part of the evaluation of an assessment. The major area of interest in this paper is how the assessment affects what happens in classrooms.

Impact on classroom practice

Alderson and Wall (1993) suggest "The notion that testing influences teaching is commonplace". This is not a new concern, with concerns raised (Vernon, 1956, Ebel, 1965) over many decades about the distorting effect on the curriculum and how teachers are induced to focus on what is measured. Alderson and Wall (1993) go on to note that the effect on teaching of tests is "a phenomenon on whose importance all seem to be agreed, but whose nature and presence have been little studied". Teaching to the test is particularly under researched in the UK context. Sturman (2003) produced a thoughtful consideration of the issues in relation to KS2 tests, but there appears to be no published work taking a similar approach involving 16-19 qualifications in the UK.

Studies in the USA (Smith, 1991) have observed the effect of assessment regimes on classroom practice over extended periods. The effect of assessment on classroom practice has also caused controversy. In the UK this has prompted a movement that focuses on the benefits of teacher assessment and advocates the use of formative assessment as a national approach. Teacher assessment may improve pedagogical practice, but due to the underdeveloped nature of our understanding of the way that teacher assessment functions in different educational environments in comparison to other forms of assessment, it cannot be viewed as a simple replacement for a robust, summative, externally assessed national examination system.

There has also been recent discussion in the UK around the 'Burden of Assessment'. Much of this centred around reducing the amount of time candidates spend in the examination room. However, the rhetoric surrounding these concerns relates to "freeing up time for deeper engagement with subjects of study" (University of Cambridge, 2004) and that time should be "given back to teachers to reintroduce the breadth and excitement of their subject" (Tomlinson, 2004). If the assessment material or mode of delivery is encouraging teachers to spend large amounts of class time focussing on the assessment material then it is increasing the burden of assessment from a teaching point of view, just as much as students spending more time in the examination hall. Considering the desire to promote breadth of teaching and deeper engagement with subjects, Alexander (2006) raises the need to guard against questions becoming “cognitively restricting rituals”.

These concerns may lead some to suggest a wholesale redesign of assessment systems in a way that totally synchronises assessment and teaching. It is interesting to consider what a complete reworking of the system could look like. At the height of the debate surrounding consequential validity a system was proposed that would change classroom practice to dovetail with assessments. This was Measurement Driven Instruction.

Measurement Driven Instruction is an approach that would be implemented across an entire administrative area promoting a highly co-ordinated and directive approach to curricular construction and assessment setting. The content of classroom instruction would be tightly proscribed and would feed directly into the end of programme tests. The approach was advocated by James Popham (1987) whose vigorous defence of an uncompromising form of Measurement Driven Instruction largely discredited the whole concept. Much of the approach had merit and a more moderate approach to Measurement Driven Instruction is outlined by Torrance (Torrance, 1993). The debate through the early 1990s was heated but died down. However, Torrance recalls that “a significant legacy of that debate [was] a recognition that assessment can shape teaching and learning, with the challenge being to make that impact positive rather than negative”. Acknowledging that classroom practice is frequently assessment driven and that it does not necessarily lead directly to adverse outcomes in the classroom provides a motivation to investigate further the positive effect that assessments can have.

In having a positive effect on the classroom, the ideas outlined by Frederiksen & Collins (1989) again are useful. The focus on cognitive skills is particularly helpful as these are the ‘things that the pupils can’t do’ as described by stakeholders (Warmington & Murphy, 2004). Much earlier Ebel (1965) focussed on the development of mental abilities as an educational goal. The idea of cognitive skills is a useful starting point in the operationalisation of this aspect of validity evaluation because it provides a link to the claims made about the assessment outcomes through grade descriptors or ‘can-do’ statements.

How classroom practice affects validity

Having considered how assessment changes classroom practice it is also necessary to explore how what happens in the classroom feeds into the validity of the assessment. The effects of assessment on classroom practice can be viewed within a taxonomy of seven approaches to test preparation developed by Koretz, McCaffrey & Hamilton, (2001) exploring the effects of classroom test preparation. The approaches can be split between meaningful gains in test scores from effective teaching, and inflated test scores from test preparation that may not be productive in a learning sense. This mirrors Messick's description of construct irrelevant variance, where Koretz's 'meaningful gains' are Messick's construct relevant variance, and Koretz's inflated gains produce construct irrelevant variance.

Koretz *et al.* distinguish between seven types of test preparation. The first three produce unambiguously meaningful gains in test scores. These are:

- Teaching more
- Working harder
- Working more effectively

A teacher working in this way enhances their students' knowledge in the subject and their ability to answer questions. The effects of this would be as described by (Massey *et al.,* 2003) regarding the gains surrounding the national tests in the UK.

Koretz *et al.* then outline less meaningful gains in test scores that can be both positive and negative.

- Reallocation
- Alignment

Reallocation occurs when "teachers report shifting instructional time to focus more on the material emphasised by an important test". This describes the action of individual teachers in deciding which parts of the curriculum to prioritise. Alignment refers to the movement of the whole school curriculum to better represent the content of test specifications.

The final two types in the taxonomy produce artificially inflated scores. Cheating clearly produces gains that are construct-irrelevant, but coaching is more ambiguous. Koretz *et al.* clarify this by suggesting that coaching that is productive in increasing the students' knowledge or skill in relation to the assessment fits into his category of reallocation. They then define coaching that only focuses on the mechanisms of the assessment as negative, producing artificially inflated gains in test scores.

- Coaching
- Cheating

Within this taxonomy, Sturman (2003) observed some elements of 'working more effectively' leading to meaningful national test gains along with successful 'reallocation' of teaching time and resources. She also observed elements of 'coaching' that led to inflated scores. Smith (1991) observed a broad spectrum of teaching strategies, many of which she classified as 'cheating'. However, she also mentions strategies for preparing students for tests that are independent of the test material. Smith describes strategies of encouraging students to do well by emphasising the importance of the tests along with teaching students methods of remaining calm and thinking clearly during the test. These she terms ‘exhortation’ and ‘stress inoculation’. They demonstrate the breadth of test preparation considerations and would result in meaningful gains. The effect of Measurement Driven Instruction would be wholesale ‘alignment’ of the school curriculum with the tests.

Coaching and cheating are clearly threats to validity, whilst reallocation and alignment are only threats to validity if important elements of curriculum have resources moved away from them such that candidates could not fulfil competency claims made about those who hold the qualifcation. Reallocation used in the right way can enhance validity. Stecher (2002) notes that “If students do not understand the test instructions or question formats, …their scores will underestimate their actual learning. removing these obstacles to performance… makes their results more valid”.

Developing a classroom survey instrument

Having established and classified these modifications to teaching and considered which are threats to validity, it is important to work towards a practical response and consider what aspects of assessments may be susceptible to construct irrelevant variance and non-meaningful gains arising from changes in classroom practice. Returning to Gipps’ (1994) earlier observation that teachers do not want to narrow their teaching it would seem productive to turn to teachers for evidence of occasions when a particular assessment is causing them to modify their classroom practice in a way they see to be counter-productive to students’ educational experience in their subject.

The concepts within consequential, curricular, systemic validities and particularly validity argument (Kane, 2006) are helpful in framing an approach to improving the quality of UK examination style assessments. However, in order to bring about change a move towards the practical application of the literature has to be made and this cannot be done through further conceptual discussion but might be done through targeted research with teachers and in classroom environments.

In a regulated environment with nearly 150 years of development, precedent and stability form a large part of how assessments are conducted. Some researchers may suggest that the web of assessment is so complex that development of formal validation requires the complete replacement of the system (Alderson & Wall, 1993, Morrow, 1986). Politically such a reworking is highly unlikely and thus incremental improvements in validity argument enable assessment agencies to evaluate, improve and defend their assessments.

Studies considering the effect of assessments in the classroom have used a range of methods. Direct long-term classroom observation on a relatively small scale (Smith, 1991), in depth questionnaires to schools (Sturman, 2003) and questionnaires to both teachers and students coupled with classroom observations (Hawkey, 2006).

Alderson and Wall (Alderson & Wall, 1993) seem to prefer directly observed evidence above that reported by the teacher. It could be suggested (Geertz, 1973) that mediation by the teacher gives a ‘truer’ view of what is happening in the classroom as the observer may not fully understand the subtleties of what is going on. This tension has existed since the early years of educational research. (Murray,1938).

Addressing the issue specifically in a classroom setting Fraser (1986) comments “Although objective indexes of directly observed behaviour in classroom settings certainly have their place in educational research, they do not tell the whole story about the complex, weighed subjective judgements made by students and others who have an important influence on learning.” He goes on to suggest that self-report is in fact a more authentic measure of what goes on in the classroom because the perceptions measured are the determinants of real behaviour. Fraser also suggests that self-report methods have advantages in that they allow access to the respondent’s thoughts aggregated from many classroom events, rather than only the small subset of events commonly observed by a researcher. Another motivation for the choice of use of a survey instrument over direct observation is that it requires fewer resources. An instrument allows the inclusion of a large number of teachers from diverse education, economic and geographical settings, which would not be as achievable with a research model based predominantly on direct classroom observation.

Large studies do not necessarily prohibit the use of direct classroom observation, but increased resources are required. Cambridge ESOL have used classroom observation in large impact and evaluation studies (Hawkey, 2006). More recently Cambridge ESOL have developed a structured method for collecting observations with the use of a video database (Hawkey, Thompson & Turner, 2006) which enables the use of multiple observations in one study without the time-consuming need for school visits by the researcher in person.

Fraser goes on to suggest that in a wider society various angles could be triangulated, specifically between observation, teacher report and pupil report. Koretz *et al.* (2001) also suggest that the only way of establishing which aspects of test score gains are meaningful is through triangulation with external measures. The external measures available (e.g. PISA (OECD, 2006) and TIMMS (NCES, 2004)) generally have wide curriculum coverage, are stable for measuring standards and their low-stakes nature negates the problems that are the subject of this paper. However, they are not a total solution as they may become less relevant over time, failing to represent adequately changes in curricular emphasis (Oates, 2007).

As the preceding review of literature demonstrates, the validity concepts surrounding the use of test items in the classroom are complex. In order to represent the closely related validity concepts faithfully in a survey instrument it is necessary to attempt to associate every item in the instrument directly with a specific concept in the literature. (Spector, 1992) concurs, suggesting that “too many scale developers spend insufficient time defining and refining the construct of interest” with DeVellis (1991) advising that “one should not overlook the importance of being well grounded in the substantive theories related to the phenomenon being studied.”

With a concept as intricate and multi-faceted as the effect of assessments on classroom practice each statement has to be traceable, justified and grounded in theory. Rust & Golombok (1999), outline a method of operationalisation through which constructs are associated with more concrete ‘manifestations’ that can form the basis of scale items. In his advice to researchers on scale construction Spector (1992), notes that “many constructs are theoretical abstractions with no known objective reality”. This is the case with validity concepts surrounding question use in the classroom. Similar issues arise as with the measurement of perceptions, as they are rather abstract traits. This is particularly relevant in relation to perceptions of external objects, in this case examination papers. The trait being measured is a respondent’s abstract concept about an object (i.e. not a person or idea). Within the educational setting (Dorman & Knightley, 2006) have developed an instrument that looks at secondary school students’ perceptions of assessment tasks.

The survey items were developed by working through the literature systematically and extracting statements that related to classroom practice to act as the major constructs in which the survey items are grounded. Whilst wording items to be neutral and not indicate to respondents whether a statement is positive or negative it may also be necessary to add items that make some measure of socially desirable response biases. This type of item would be taken from a previously validated instrument and enable an assessment of how susceptible the respondents are to giving responses that represent themselves in what they perceive to be a positive way. Responses to each item will be recorded on a standard Likert-type four category opinion scale. Rust & Golombok, (1999) advise that response scales should not have a middle option to avoid the tendency of respondents towards indecisiveness on difficult items.

This results in the development of a list of possible survey items that could be used with teachers to identify the questions that are a potential threat to the validity of the overall assessment because of their perceived susceptibility to ‘teaching to the test’.

These prospective survey items are shown below along with the literature from which they are derived.

This question could be used to help students revise.
This question is less useful for revision than similar questions on the same topic.
This question would be useful as a teaching aid.
The question would cause me to teach this concept differently in future years.

Educational benefits due to enhanced systemic validity result from *“evolutions in the form of and content of instruction engendered by use of the tests”.*
Questions should *“serve as a beacon to guide future learning”.*
Frederiksen and Collins (Frederiksen & Collins 1989).

Students have to learn how this question style works to gain full marks.

“an over-emphasis on [summative assessment] leads to a highly instrumentalised and surface approach to learning”

The question credits the ability to read critically.
The question credits the ability to communicate ideas.
The question encourages the ability to read critically.
The question encourages students who communicate ideas well.
This question is more about knowing the material than communicating ideas.
Able candidates who hadn’t seen this type of question before would do well on it.

“the ability to read critically [and] to communicate ideas in writing” is a desired outcome.
There are *“certain core ideas that need to be understood”.*

This question tests ideas that most students will know.
Most of my students understand the ideas needed for this question.
This question rewards original answers.
This question encourages original answers.
A range of skills are required to answer this question.
Students would enjoy answering this question.

“current assessments are counter productive to the desirable skills of engaging critically, taking intellectual risks and using a range of skills”
The Nuffield review (Wilde, Wright, Hayward, Johnson & Skerrett, 2006).

This question would be enjoyed by inquisitive students.

This question might frustrate more able students.

More able students would find this question boring.

This question could be approached in a number of ways.

Items identifying questions that expand the capacity to learn (Claxton, 2006).

This question format will be familiar to students.

Students may be surprised by the format of this question.

I would use exercises similar to this question as teaching material.

To be successful students need to be familiar with this question style.

Items identifying threats to validity derived from the work of Stecher (2002).

In the trial of the instrument these items will be administered alongside a set of past examination questions, with responses required on all the items for each examination question in turn. It is intended to administer the items above to teachers and examiners in the near future and then to produce a scale of ten to fifteen items that could be used in a measurement tool for evaluating examination questions.

Conclusion

This approach begins to respond to a number of contemporary calls for action. In the preliminary report of the Nuffield Review, (Wilde *et al.,* 2006) identify seven key emerging issues. One of these is that narrow accountability due to high stakes testing leads to "spoon feeding rather than the fostering of independence and critical engagement with the subject material", concluding that "discussion needs to be fed back effectively into procedures for qualification and curriculum design". Examination questions frequently feed into informal curriculum design (what actually happens in classrooms) only too readily in their use as teaching aids for exam preparation. A deeper understanding of the way in which different question types encourage a teacher to align their class curriculum, reallocate teaching time, or even directly coach for high-stakes assessments, would enable those preparing questions to create them so as to promote good classroom practice and would help teachers to use past questions in a positive and educationally valuable way.

Ultimately, for assessments to have value to society and the economy, they need to result in candidates developing the skills that the assessment designers intended. Gipps (1994) argues that "what we need to know is that students have been taught, not the items in the test, but the skills and knowledge measured by the test". McGaw (2006) notes that high-stakes assessment encourages a focus on the elements of a curriculum that appear in the tests, which he suggests "makes it all the more important that the assessment induces focus on what is important".

Messick (1996) suggests that assessments may not of themselves be able to create positive classroom practice, but if they can avoid inducing negative practice then the assessment is improved. The instrument in development described here, in providing greater understanding of the effect of an assessment in the classroom may help to improve the validity of assessment practice and result in a higher quality assessment.

References

Alderson, J.C. & Wall, D., (1993), Does Washback Exist? *Applied Linguistics*, **14,** 115-129.

Alexander, R., (2006), *Towards dialogic teaching: rethinking classroom talk.* Cambridge: Dialogos.

Claxton, G., (2006), *Expanding the Capacity to Learn: A new end for education?* Keynote address given at the British Educational Research Association Conference, University of Warwick, September 2006.

DeVellis, R. F., (1991), *Scale Development: Theory and Applications.* Applied Social Research Methods Series. California: SAGE.

Dorman, J.P. & Knightley, W.M., (2006), Development and validation of an instrument to assess secondary school students' perceptions of assessment tasks. *Educational Studies*, **32,** (1), 47-58.

Ebel, R L., (1965), *Measuring educational achievement*. Englewood Cliffs, NJ: Prentice-Hall.

Fraser B J., (1986), *Classroom Environment.* Beckenham: Croom Helm.

Frederiksen, J.R. & Collins, A., (1989), A Systems Approach to Educational Testing. *Educational Researcher*, **18,** (9), 27-32.

Geertz, C., (1973), *The Interpretation of Cultures.* London: Fontana.

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

Hamp-Lyons, L., (1997), Washback, impact and validity: ethical concerns, *Language Testing,* **14,** (3), 295-303

Hawkey, R., (2006), *Impact Theory and Practice: Studies of the IELTS test and Progetto Lingue 2000,* Studies in Language Testing 24. Series Editors, Milanovic, M and Weir, C. Cambridge: Cambridge University Press.

Hawkey, R., Thompson, S. & Turner, R., (2006), Developing a classroom video database for test washback research. *Research Notes*, **26**, 5-9.

Kane, M.T., (2006), *Validation*, in Brennan, R.L. *Educational Measurement,* 4th edition. Westport, CT: Praeger.

Koretz, D. M., McCaffrey, D. F. & Hamilton, L. S., (2001), *Toward a Framework for Validating Gains Under High-Stakes Conditions*. Centre for the Study of Evaluation: Technical Report 551. Los Angeles, CA: University of California.

Massey, A. J., Green, S., Dexter, T. & Hamnett, L., (2003), *Comparability of national tests over time: KS1, KS2 and KS3 standards between 1996 and 2001*. Qualifications and Curriculum Authority.

McClung, M.S., (1978), Are competency testing programmes fair? cited in Wood, R. (1991) *Assessment and Testing : A survey of research,* Cambridge: University of Cambridge Local Examinations Syndicate.

McGaw, B., (2006), *Assessment Fit for Purpose*. Keynote address given at the 32nd Annual Conference of the International Association for Educational Assessment, Singapore

Mehrens, W.A., (1997), The Consequences of Consequential Validity. *Educational Measurement: Issues and Practice*, **16,** (2), 16-18.

Messick, S., (1993), *Validity*, in Linn, R.L. *Educational Measurement,* 3rd edition, Pheonix, AZ: American Council on Education and The Oryx Press.

Messick, S., (1995), Standards of Validity and the Validity of Standards in Performance Assessment. *Educational Measurement: Issues and Practice*, **14,** (4), 5-8.

Messick, S., (1996), *Validity and washback in language testing*. Princeton, NJ:Educational Testing Service.

Moss, P.A.., (1992), Shifting Conceptions of Validity in Educational Measurement: Implications for Performance Assessment. *Review of Educational Research*, **62,** (3), 229-258.

Murray, H. A., (1938), *Explorations in Personality*. New York: Oxford University Press.

NCES, (2004), *Highlights from the Trends in International Mathematics and Science Study (TIMSS) 2003*, Washington DC: NCES

Oates, T., (2007), *Viable approaches to measuring the performance of education systems at national level*, Working paper.

OECD, (2006), Assessing Scientific, Reading and Mathematical literacy: A framework for PISA 2006, OECD.

OCR, (2007a), *English Language GCE: Report on the Units, January 2007*, Cambridge: OCR

OCR, (2007b), *Mathematics GCE: Report on the Units January 2007*, Cambridge: OCR

Popham, W.J., (1987), The Merits of Measurement-Driven Instruction. *Phi Delta Kappan*, **68**, (9), 679-682.

Rust, J. & Golombok, S., (1999), *Modern Psychometrics: The Science of Psychological Assessment*. London: Routledge.

Shepard, L.A., (1997), The Centrality of Test Use and Consequences for Test Validity. *Educational Measurement: Issues and Practice*, **16,** (2), 5-8.

Smith, M.L., (1991), Meanings of Test Preparation. *American Educational Research Journal*, **28,** (3), 521-542.

Spector, P. E., (1992), *Summated Rating Scale Construction*. California: Sage.

Stecher, B.M., (2002), *Consequences of large-scale, high-stakes testing on school and classroom practice*. in Hamilton, L.S., Stecher, B.M. & Klein, S.P., (2002), *Making Sense of Test-Based Accountability in Education.* RAND Corporation.

Sturman, L., (2003), Teaching to the test: science or intuition? *Educational Research*, **45,** (3), 261-273.

The Advisory Committee on Mathematics Education, (2002), *Response to the DfES consultation document '14-19: extending opportunities, raising standards'*. downloaded 09/08/07 from www.royalsoc.ac.uk/acme/acme_gp.pdf

Tomlinson, M., (2004), *14-19 Curriculum and Qualifications Reform*. London, DfES.

Torrance, H., (1993), Combining Measurement-Driven Instruction With Authentic Assessment: Some Initial Observations Of National Assessment in England and Wales. *Educational Evaluation and Policy Analysis*, **15,** (1), 81-90.

Torrance, H., (1995), The role of assessment in educational reform, In Torrance, H. (ed.) *Evaluating Authentic Assessment,* Buckingham: Open University Press.

University of Cambridge, (2004), Response to the Interim Report of the Working Group on 14-19 Reform, downloaded 06/08/07 from http://www.cam.ac.uk/admissions/undergraduate/responses/reform1419b.doc

Vernon, P.E., (1956), *The Measurement of Abilities,* London: University of London Press.

Warmington, P. & Murphy, R., (2004), Could do better? Media depictions of UK educational assessment results. *Journal of Education Policy*, **19,** (3), 285-299.

Weiss, C., (1998), *Evaluation,* New Jersey: Prentice Hall.

Wilde, S., Wright, S., Hayward, G., Johnson, J. & Skerrett, R., (2006), *Nuffield Review Higher Education Focus Groups Preliminary Report*, The Nuffield Review of 14-19 Education & Training.