The validity of tutor assessed Independent Research Reports contributing to a pre-university qualification

Jackie Greatorex, Research Division, Cambridge Assessment

Stuart Shaw, Cambridge International Examinations

Corresponding author:
Email Greatorex.j@cambridgeassessment.org.uk
Telephone 01223 553835

Abstract

This research considers the validity of tutor assessed, pre-university independent research reports. Evidence of construct relevance in tutors' interpretations of the levels awarded to the candidates' research process was investigated. This includes designing, planning, managing and conducting their own research project using techniques and methods appropriate to the subject discipline. The research was conducted in the context of the Cambridge Pre-U Global Perspectives and Independent Research qualification (the GPR), a pre-university qualification for 16-19 year olds which is designed to equip students with the skills required to make a success of their university studies. Tutors' justifications for the levels they gave candidates were considered. In the first of two studies (Study 1), tutor justifications were qualitatively analysed for specific tutor behaviours that might highlight tutors interpreting levels in a construct irrelevant way. In the second study (Study 2), external moderators (EMs) rated the justifications according to the extent to which they reflected the intended constructs. Study 1 showed little evidence of construct irrelevance and Study 2 provided strong evidence of construct relevance in tutors' interpretation of the levels they awarded candidates for the research process.

Keywords

Validity, internal assessment, moderation, examinations, construct relevance.

ISSN: 1758-2199

Introduction

Candidates preparing for higher level education in the UK and internationally are sometimes given the challenge of conducting a piece of independent research which may entail designing, planning and managing a research project; collecting and analysing information; evaluating and making reasoned judgements; and communicating findings and conclusions in a written report. It also enables candidates to develop a range of generic research skills including practical and analytical research, higher order thinking, interpretation and time management. When applying to university, candidates can use their research reports to demonstrate motivation for their intended course of study and to differentiate themselves from competing applicants.

Following the recommendations of Tomlinson (2004), some schools now offer candidates the opportunity to conduct research projects which are formally assessed and contribute to a formal qualification. For some of these qualifications, the candidates' research reports are assessed by their own tutors. The tutors' marks are then moderated by EMs who are contracted by the examination board administering the qualification. The Cambridge Pre-U Independent Research Report, administered by Cambridge International Examinations, utilises this assessment approach, as do the extended project qualifications administered by the AQA, OCR, and Edexcel examination boards.

The research evaluated evidence of construct relevance in tutors' justifications for the levels they award candidates' research process. Before reporting the research an introduction to construct relevance and the context of the research - Cambridge Pre-U Global Perspectives and Independent Research qualification - is provided. A glossary of terms used in this paper is provided below.

ISSN: 1758-2199

Glossary

Cambridge Nationals	Formerly OCR Nationals. Vocationally related qualifications that take a practical approach and are industry related. Available for teaching from September 2012.
DfE	Department for Education.
EM	External moderator. Assessor contracted by Cambridge International Examinations to check the tutors' assessment decisions and compare standards between centres. EMs recommend how tutors' assessments should be scaled (adjusted) if necessary.
Extended project qualification	A pre-university qualification designed to provide an opportunity for candidates to: • choose and design an extended piece of work • develop as critical and independent learners • develop and apply decision making and problem-solving skills • extend their research, critical thinking, analysis, synthesis, evaluation and presentations skills.
IM	Internal moderator. Tutor who leads and quality assures the assessing within a centre.
Inset	In service training for teachers.
Key Stage 3	Three years of schooling in government funded schools in England and Wales. That is Year 7, Year 8 and Year 9 when pupils are aged 11 to 14.
OCR Nationals	Examination-free, vocationally related qualifications that took a practical approach and were industry related. They were primarily designed for 14-19 year olds. The last certificates were issued in 2012.
Ofqual	National regulator of government funded qualifications, excluding university qualifications.
QCA	Qualification and Curriculum Authority. Predecessor of Ofqual.
SCAAT	School and College Achievement and Attainment Tables. Tables ranking schools and colleges in England by examination results.
Standardisation meeting	Meeting of tutors to ensure assessment decisions are consistent.
Tutor	Pre-U teacher in a centre.
UCAS	Universities and Colleges Admissions Service. An organisation that manages applications to higher education courses at universities and colleges.

Construct relevance

Given the fundamental aim of ensuring that educational assessments are valid, it is important for examination boards to be able to provide evidence of

ISSN: 1758-2199

the validity of their assessments. Ensuring that the intended constructs are being measured by an assessment is important to validity. The aim of any educational assessment, therefore, is to limit assessment to only attributes relevant to the construct the assessment intended to measure and to curb the influence of attributes irrelevant to the intended construct (Bracken, 2000). An assessment must be a good measure of the attribute it is interpreted to assess (Messick, 1975).

It is important to make sure that the constructs elicited are precisely those intended and that scores are not contaminated by other irrelevant constructs. Messick (1989) highlighted two potential threats to validity: (a) construct under-representation (i.e. the test is too narrow in focus and fails to include important elements of the construct of interest), and (b) construct irrelevant variance (i.e. a type of systematic measurement error where test score variance is due to factors other than the construct of interest, such as background/cultural knowledge, or unreliable scoring). A consequence of construct under-representation is that the test results are unlikely to reveal a candidate's true abilities within the construct. Construct irrelevant variance suggests that the assessment measures too many variables, some of which are irrelevant to the intended construct.

Construct relevance and irrelevance are not properties of assessments, they are the properties of the meaning of marks, grades, levels and so on (Messick, 1995, Shepard, 1993, Messick, 1989). A claim about construct relevance is an evaluative evidence based judgement about the extent to which marks/grades are interpreted as the skills and knowledge the assessment was intended to measure. The evidence includes candidates' performance and cognition, statistical analysis of marks (facility values/item response theory), the assessment tasks and the tutors' interpretation/application of the mark scheme. The judgements and their basis need to be open to scrutiny in line with validity theory and assessment validation (Shepard, 1993, Xi, 2008, Stobart, 2009, Crisp and Shaw, 2010, Crisp, 2010, Crisp and Shaw, 2011, Shaw, Crisp and Johnson, 2011, Messick, 1989, Kane, 2009, Sireci, 2007, Sireci, 2009) as well as for the purposes of transparency. Examination board guidance and research state that evidence based judgement of construct relevance is a thread running throughout assessment development, revision and use (Berk, Lohman and Cassata, 2001, Quinlan, Higgins and Wolff, 2009, Oates, 2009).

There are empirical studies focusing on construct relevance. Key research pertinent to the UK is presented below.

Candidates' responses to tasks (examination questions), marks gained on tasks and tasks themselves were studied to identify construct relevant and construct irrelevant sources of difficulty (Pollitt, Entwistle, Hutchinson and de Luca, 1985, Hughes and Fisher-Hoch, 1997). The results were used to make recommendations for setting tasks that focused on assessing the intended construct. Later Pollitt and Ahmed (1999) argued that the construct being assessed is a psychological process or a combination of mental activities

ISSN: 1758-2199

required to perform at a certain level of proficiency in responding to the assessment task. Therefore they developed a task response model of six steps:

- Learning the subject/ field.
- Understanding the task.
- Accessing relevant aspects of memory.
- Re-interpreting stored knowledge to match the assessment task.
- Generating a response to the task.
- Providing a response to be assessed.

The validity of the marks/grades is a function of how well the candidates' mental activities corresponded with the intended construct, that is whether candidates' mental activities are construct relevant. This line of reasoning runs through other research such as Daneman and Hannon (2001) and Threlfall, Nelson and Walker (2007).

Some researchers challenged the validity of the results of some reading comprehension assessments claiming that candidates could allegedly gain better-than-chance marks even if they did not read and comprehend the passage on which the assessment was based. To investigate the claims, Daneman and Hannon (2001) experimented with which assessment-taking strategies, working memory capacity and reading strategies were successful in gaining certain assessment results. They found that the assessment appeared to be tapping reading comprehension processes as long as the candidates read at least some of the passage, that is, they found evidence of construct relevance.

Threlfall, Nelson and Walker (2007) investigated the sources of difficulty in the Key Stage 3 ICT assessments. They argued that if the sources of difficulty related to the intended construct (ICT capability) then they were construct relevant. However, if the source of difficulty did not relate to ICT capability it was designated construct irrelevant. They used the assessment tasks, pupils' performance and pupils' cognition to evidence construct relevance. The sources of difficulty related to the task and software (e.g. insufficiently explicit task instructions), as well as how prepared the pupils were for the assessment (e.g. a pupil spending time achieving an aesthetically pleasing response rather than focusing on the task). They found that there were some construct relevant (and irrelevant) sources of difficulty.

Greatorex and Shaw (2012) developed a qualitative coding system (Figure 1below) to classify tutors' comments which evidenced a tutor attending to an irrelevant characteristic of the candidate or performance. If a substantial number of comments evidenced tutors attending to irrelevant characteristics then this would arguably be evidence of construct irrelevance. They found that only 5 out of the 150 comments evidenced tutors attending to an irrelevant characteristic of the candidate/performance and therefore there was no real evidence of construct irrelevance.

ISSN: 1758-2199

Figure 1: Assessor behaviours found in the literature which underpinned coding categories

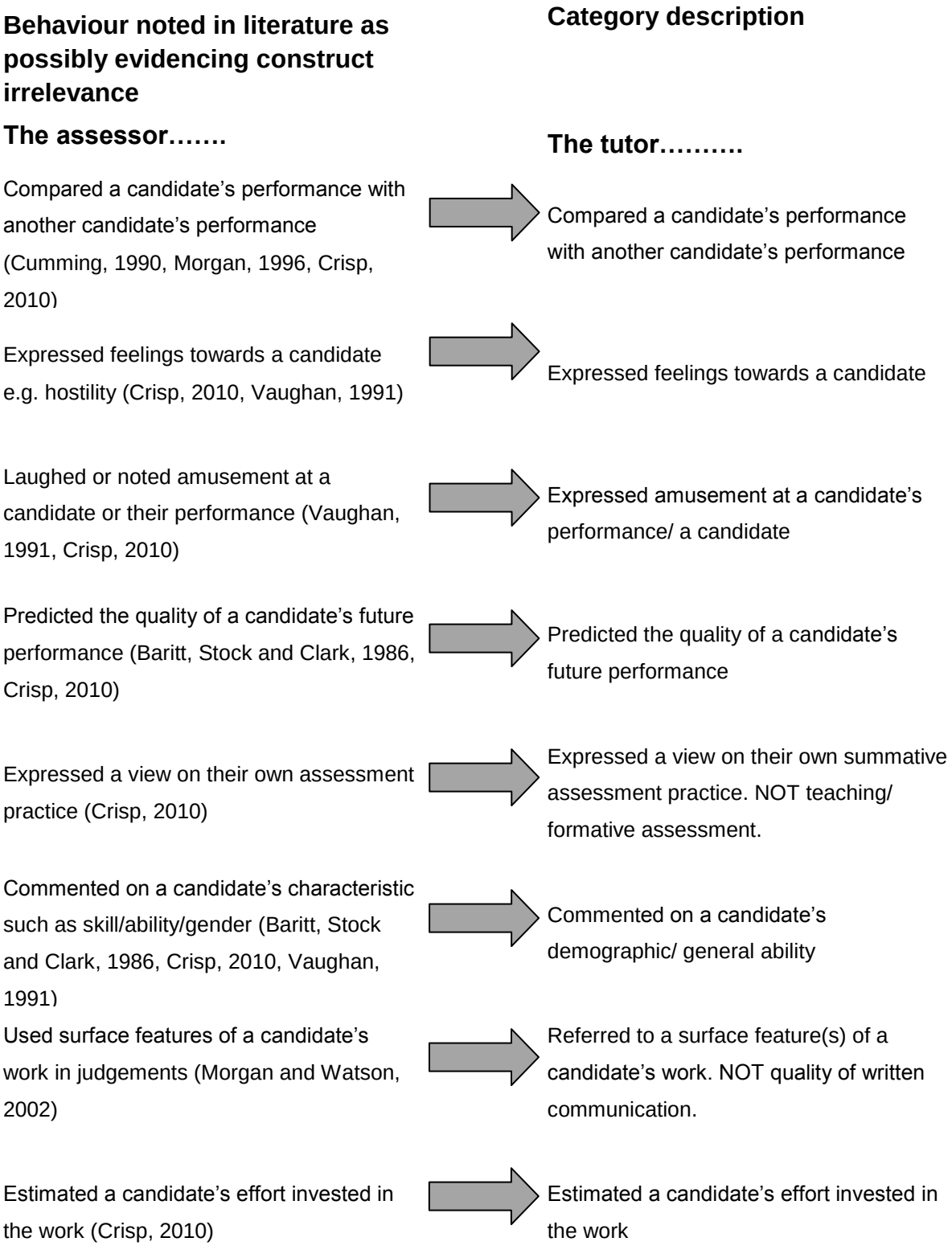

Figure 1 is adapted from Greatorex and Shaw (2012) page 40 with permission from the Editor of Research Matters: A Cambridge Assessment Publication.

ISSN: 1758-2199

The Cambridge Pre-U Global Perspectives and Independent Research

Cambridge Pre-U Global Perspectives and Independent Research is an international post-16 qualification intended to prepare candidates for undergraduate studies by stretching and challenging them. It is provided by Cambridge International Examinations. The first cohort of Cambridge Pre-U candidates completed their courses in the summer of 2010.

Typically Cambridge Pre-U candidates study three Principal Subjects over a two-year period (or alternatively, a combination of Principal Subjects and A Levels). In addition to this, to obtain the Cambridge Pre-U Diploma, they must complete the Cambridge Pre-U's course in Global Perspectives and Independent Research (GPR). GPR is known as the core of the Cambridge Pre-U Diploma. It comprises two components: the Global Perspectives course (GP), and the Independent Research Report (IRR) which may be up to 5000 words long.

The IRR constructs to be assessed are given (either explicitly or implicitly) in:

- The mark scheme (http://www.cie.org.uk/qualifications/academic/uppersec/preu/subjects/subject/preusubject?assdef_id=1018).
- The Tutor Monitoring Form – TMF.
- Guidance to centres including standardisation meetings and visits to centres. (As part of the support for centres as they undertake the IRR, Cambridge International Examinations offers a visit by a Verifier to help centres ensure their internal procedures for setting up the supervisory/tutorial system and managing the internal standardisation of tutors and awarding marks and levels.)
- In service training (Inset).

The IRR is assessed by a tutor at the candidate's centre. The assessment comprises the report, a five to ten minute terminal interview (viva) to authenticate that the candidate did the work and the tutor's observations, experience and records of the candidate's progress in developing and producing the research report.

Samples of marking are centre moderated by an internal moderator (IM) and externally moderated by Cambridge International Examinations. External moderation checks the marking of the report. It should be noted that part of the mark scheme is for internally assessing and internally moderating each candidate's 'Knowledge and understanding of the research process' (AO1) as evidenced by their tutor's observations, experiences and records. That is AO1 is not externally moderated. This situation arises as EMs do not have the tutor's and IM's experience of the candidate to judge AO1.

The TMF allows tutors to record a level for each of three aspects of AO1, namely -identifying the research question, developing the question and planning and carrying out independent research. They also record a level and

ISSN: 1758-2199

a justification for each level. The TMF levels and justifications facilitate holistic assessment decisions about the level and number of marks to award the candidate for AO1. Completing the form is voluntary - it is not a requirement of tutors. The TMFs are part of the script (candidate's report).

The TMF justifications communicate the meaning the tutors' assigned to the levels they credited to candidates. Consequently, evidence pertaining to construct relevance was gathered as part of general IRR practice. The opportunity for post hoc research about evidence of construct relevance in the context of AO1 was recognised and used. The present studies explore the practical application of one of the five different generic assessment objectives (AO1) which comprises the IRR mark scheme and which can only be assessed in the context of the classroom, by candidates' own tutors. The studies form part of a wider on-going research programme supporting the Cambridge Pre-U (Shaw and Suto, 2010, Suto and Shaw, 2010).

Here two studies are reported. In Study 1 justifications were qualitatively analysed for tutor behaviours that might indicate construct irrelevance. In Study 2 the justifications were rated by EMs regarding whether they indicated construct relevance. The scope of the research is construct (ir)relevance in tutors' use of AO1 and does not provide a full validity argument and evidence for the IRR.

Study 1: Tutor behaviours

Method

Data

The justifications were sourced from all available TMFs. The justifications were generated from 61 of the 92 IRR candidates. That is, the justifications were from seven of the 11 centres that entered candidates. This provided a qualitative data set of 183 justifications. The names of centres, tutors and IMs were replaced with codes. Candidates' names were replaced with case numbers unless they occurred in the justifications when they were replaced with "CANDNAME". The justifications were also anonymised regarding research question (topic), place names and profession(s), in this report they are represented by "…".

Qualitative coding

An established qualitative coding system (Greatorex and Shaw, 2012) was used to analyse each justification for the presence/absence of each tutor behaviour. A coder read the justification and coded each one according to the categories. A second coder blind coded the same comments. Any disagreements were passed over to a third coder for adjudication.

ISSN: 1758-2199

Results

A tutor behaviour was present in only 15 out of 183 justifications (Table 1 below). The scarcity of the tutor behaviours was evidence that the tutors' interpretation of levels did not manifest construct irrelevance.
The justifications contained only four types of behaviours, i.e. tutors:

- Expressed feelings towards a candidate.
- Predicted the quality of a candidate's future performance.
- Commented on a candidate demographic/ general ability.
- Estimated a candidate's effort invested in the work.

There were eight justifications for which the presence or absence of a tutor behaviour remained unresolved, that is the coders did not agree. This was perhaps due to ambiguity in the justifications or categories.

Table 1: Frequency of justifications in a category

Category **The tutor….**	**Examples from IRR data**	**AO1 1 Identifying the research question**	**AO1 2 Developing the question**	**AO1 3 Planning and carrying out**	**Unresolved**
Compared a candidate's performance with another candidate's performance	-	0	0	0	0
Expressed feelings towards a candidate	CANDNAME was an excellent student who really owned the process; the questions were hers; she sourced relevant materials and reading; she read them thoroughly; and considered them thoughtfully. She needed no help in staying on task, staying focused and meeting	1	0	3	4

ISSN: 1758-2199

Category The tutor….	Examples from IRR data	AO1 1 Identifying the research question	AO1 2 Developing the question	AO1 3 Planning and carrying out	Unresolved
	deadlines. Her design, her research and the production of her study was meticulous. In discussion with her she was always positive, considered, very well informed, very clear as to her direction. She was an absolute delight to work with.				
Expressed amusement at a candidate's performance/ a candidate	-	0	0	0	0
Predicted the quality of a candidate's future performance	CANDNAME needed help with structuring her report, and in planning a structure. She did locate the resources largely on her own. Without this support I believe she would not have produced the study and she needed encouragement to complete the course.	0	0	1	0
Expressed a view on their own summative	-	0	0	0	0

ISSN: 1758-2199

Category The tutor….	Examples from IRR data	AO1 1 Identifying the research question	AO1 2 Developing the question	AO1 3 Planning and carrying out	Unresolved
assessment practice. NOT teaching/ formative assessment.					
Commented on a candidate demographic/ general ability	The candidate was very well prepared for the regular meetings held. He showed considerable initiative in identifying and then ordering appropriate resources. He developed into a very strong independent learner.	0	0	1	1
Referred to a surface feature(s) of a candidate's work. NOT quality of written communication.	The candidate showed clear initiative from the start. The candidate chose the topic in question because she had ….. The candidate developed the focus for study and question in conjunction with her plans for travel.	1	0	0	0
Estimated a candidate's effort invested in the work	CANDNAME continued to read and write in a steady way on her own resources, but	1	0	7	3

ISSN: 1758-2199

Category The tutor….	Examples from IRR data	AO1 1 Identifying the research question	AO1 2 Developing the question	AO1 3 Planning and carrying out	Unresolved
	did require some encouragement at several meetings to seek out a wider range of texts and to develop the research in a more explorative way.				

A limitation of Study 1 is that it does not evaluate construct relevance, it merely shows a lack of construct irrelevance (which is not the same as evidence of high construct relevance). This limitation was tackled in Study 2, which aims to evaluate the alignment between the intended construct and the constructs evidenced in justifications.

Study 2: EMs' judgements

Participants

The two participants were EMs and experts in the intended construct.

Materials

The data collection instrument constituted tables containing a column of justifications and other columns for each point on a four point rating scale. Each row contained a justification and radio buttons for the EM to record their judgement about the justification. The EMs were instructed to indicate to what extent each justification fitted with their view of AO1. Sections from the form including prompts and the rating scale are provided as Table 2, Table 3 and Table 4.

Procedure

The EMs were sent the instrument in hard copy to be completed remotely. They were asked to rate each justification on the four point scale. The responses were keyed into a database. Some ratings were not accurately recorded in the radio buttons. A research assistant judged which response option was chosen, where possible.

Please indicate to what extent each justification below (column 1) fits with your view of AO1

Table 2

ISSN: 1758-2199

1. Identify the research question Tutor justification (from Independent Research Report Monitoring Form for Tutors)	Greatly reflects AO1	Adequately reflects AO1	Inadequately reflects AO1	Bears no relation to AO1
The candidate had already got a long way into the report topic when I first met him.	O	O	O	O

Please indicate to what extent each justification below (column 1) fits with your view of AO1

Table 3

2. Developing the question Tutor justification (from Independent Research Report Monitoring Form for Tutors)	Greatly reflects AO1	Adequately reflects AO1	Inadequately reflects AO1	Bears no relation to AO1
Candidate certainly showed ability to think about implications of the question and its wording. Dialogue certainly constructive, clearly indicative of initiative.	O	O	O	O

ISSN: 1758-2199

Please indicate to what extent each justification below (column 1) fits with your view of AO1

Table 4

3. Planning and carrying out **Tutor justification (from Independent Research Report Monitoring Form for Tutors)**	**Greatly reflects AO1**	**Adequately reflects AO1**	**Inadequately reflects AO1**	**Bears no relation to AO1**
He required some direction in terms of structuring and some help was needed directing him to theories.	O	O	O	O

ISSN: 1758-2199

Results

Frequencies of ratings (Table 5) were used to investigate whether the justifications reflected AO1. Most of the ratings (98%) were "greatly reflects" or "adequately reflects" and this was indicative of the justifications reflecting the intended construct and therefore evidence of considerable construct relevance in the tutors' interpretation of levels.

Table 5 Frequency of ratings given by the EMs for justifications

Variable	Greatly reflects AO1	Adequately reflects AO1	Inadequately reflects AO1	Bears no relation to AO1	Missing data
Rating given by the EMs for AO1 1 Identifying the research question Tutors' justifications	99	18	5	0	0
Rating given by the EMs for AO1 2 Developing the question Tutors' justifications	35	68	12	5	2
Rating given by the EMs for AO1 3 Planning and carrying out Tutors' justifications	16	85	21	0	0

A limitation of Study 2 is that the rating scale refers to "reflecting the AO" rather than construct relevance per se. Experience shows that EMs consider construct relevance but do not tend to use technical language such as construct, construct relevance etc. Therefore the scale used (Table 5) was probably more accessible to the EMs than a scale about constructs.

ISSN: 1758-2199

Discussion

Study 1 provides strong evidence of a lack of construct irrelevance and Study 2 offers solid evidence of construct relevance in assessment of 'Knowledge and understanding of the research process' (AO1) as evidenced by tutor's observations, experiences and records. Also the findings provide evidence that the procedures in place to communicate the intended construct are successful. These findings are important to stakeholders – candidates, centres, higher education and employers.

Furthermore, the findings are of theoretical interest. Previous research about other qualifications/assessments showed that assessor behaviours occurred which possibly highlighted some construct irrelevance (Morgan, 1996, Morgan and Watson, 2002, Vaughan, 1991, Crisp, 2010, Barritt, Stock and Clark, 1986, Cumming, 1990). The presence of the behaviours is reported in the literature, but the frequency of the behaviours is not always given. Given evidence of these behaviours in the literature and a lack of these behaviours in using AO1, the findings of the present study appear somewhat unusual.

Established research shows that tutors from different subject specialisms may have different interpretations of assessment terms or candidates' work which might lead to a lack of a shared understanding of the construct or inconsistent application of the assessment criteria (Nadas, Suto and Grayson, 2012, Sadler, 1989, Shay, 2006, Johnson, 2008). This finding holds even when standardisation processes (or similar) are in place. The present research assumed that the tutors had a variety of subject specialisms as candidates could choose to research any question beyond any of the Pre-U subject syllabuses. The literature suggests that the tutors would interpret the criteria in a variety of ways. However, the Study 2 results did not follow this trend. The analysis showed that the tutors interpreted AO1 in a similar way to the EMs. It might be that the standardisation processes and other forms of centre support are useful in facilitating construct relevance. Perhaps AO1 also focuses on skills which are very general to research in many subjects and therefore this aided a common interpretation.

The findings are particularly interesting in the current educational context in England for several reasons.

Firstly, they provide confidence in internal assessment when there has been a lack of confidence in internal assessment in England. This can be illustrated by recent events regarding GCSE coursework (a form of internal assessment). QCA (2005) found that the level of control in GCSE internal assessment should increase to give the public more confidence in the qualifications. The following findings were reported in QCA commissioned research which collected tutors' views regarding internal assessment:

- 60% thought it would be easy for them or their colleagues if candidates conducted coursework under supervised conditions
- 14% wanted to "do away with/scrap coursework / prefer exams", 10% wanted a "reduction in the amount of coursework/assignments required",

ISSN: 1758-2199

> 13% wanted "coursework to be done under supervised/examination conditions" and 7% wanted "coursework to be done in allocated time".

For further details see Ipsos MORI (2006). The views were reflected in the introduction of controlled assessment (rather than coursework) to GCSE. Controlled assessment is more regulated than coursework and is also more tightly defined and managed (Isaacs, 2010). There are rules for how each subject's assessment is set and conditions under which it is marked. However, once the changes to controlled assessment were introduced, it was believed by some that they resulted in a loss of teaching and learning time and a reduction in learning (Crisp and Green, 2012, Ipsos MORI, 2011).

Second, it is interesting to note that at a time when there seems to be pressure to increase external assessment, the Study 1 and 2 findings are favourable towards internal assessment. Vocational qualifications typically had little or no external assessment (Isaacs, 2010). Subsequently, the Government decided that vocational qualifications must meet several criteria for the qualification holders' results to be in School and College Attainment and Achievement Tables (SCAAT), and that external assessment should contribute a minimum of 20% towards the candidate's final grade (DfE, 2011). It is likely that examination boards have more chance of selling a qualification to a school if the qualification's results are in SCAAT. This is because SCAAT content can contribute to parents choosing a particular school for their child and forms of accountability. Perhaps the Government's decision encouraged examination boards to increase the amount of external assessment in vocational qualifications so that their results would remain in SCAAT. Indeed, OCR Nationals were examination free and since the Government's decision Cambridge Nationals (formerly OCR Nationals) have at least 20% external assessment.

Third, the research provides evidence of construct relevance and a lack of construct irrelevance when assessing project work using a viva and the tutors' observation, experience, and records. In contrast, many general qualifications assess using only an artefact the candidate produced, e.g. most A Level assessment is via candidates providing written responses to examination question papers. Perhaps the research findings somewhat corroborate Stacey's (Chief Executive of Ofqual) suggestion that there should be an increase in project work and alternative forms of assessment such as oral examinations in A Levels (Clark, 2012).

Conclusion

This research provides strong evidence of a lack of construct irrelevance and offers solid evidence of construct relevance in assessing the 'Knowledge and understanding of the research process' (AO1), and therefore no threats to validity were identified. This adds to the body of research supporting internal assessment of candidates' IRR performance (Suto and Shaw, 2010, Shaw

ISSN: 1758-2199

and Suto, 2010), and the validity of internal assessment more generally. The findings suggest that AO1 facilitates valid assessment.

The results lend some support to recent calls for more project work and oral examinations, and to those who advocate internal assessment despite political pressures which might increase external assessment in some areas of education.

References

Barritt, L., Stock, P. L. & Clark, F. (1986). Researching practice: evaluating student essays. *College Composition and Communication*, 37(3), 315-327.

Berk, E. J. V., Lohman, D. F. & Cassata, J. C. (2001). What does a verbal test measure? A new approach to understanding sources of item difficulty. *Annual Meeting of the American Educational Research Association*. Seattle, WA.

Bracken, B. A. (2000). Maximizing construct relevant assessment: the optimal preschool testing situation. *In:* Bracken, B. A. (ed.) *The Psychoeducational Assessment of Preschool Children*. 3rd ed. Boston: Allyn & Bacon.

Clark, L. (2012). Pupils 'need more coursework and multiple choice questions at A-level', says exam watchdog. *Mail Online*, 2nd February.

Crisp, V. (2010). *The judgement processes involved in assessing GCSE coursework.* PhD, University of London.

Crisp, V. & Green, S. (2012). The effects of the change from coursework to controlled assessment in GCSEs. *British Educational Research Association* Manchester, UK.

Crisp, V. & Shaw, S. D. (2010). How hard can it be? Issues and challenges in the development of a validation method for traditional written examinations. *International Association for Educational Assessment Annual Conference.* Bangkok, Thailand.

Crisp, V. & Shaw, S. D. (2011). Applying methods to evaluate construct validity in the context of A level assessment. *Educational Studies*, 38(2), 209-222.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, 7(1), 31-51.

Daneman, M. & Hannon, B. (2001). Using working memory theory to investigate the construct validity of multiple choice reading comprehension tests such as the SAT. *Journal of Experimental Psychology*, 130(2), 208-223.

ISSN: 1758-2199

DfE. (2011). Qualifications for 14-16 Year Olds and Performance Tables. Technical guidance for Awarding Organisations. Available: http://media.education.gov.uk/assets/files/pdf/c/consultation%20response%20on%20qualifications%20for%2014-16-year-olds%20and%20performance%20tables.pdf [Accessed 28 October 2011].

Greatorex, J. & Shaw, S. D. (2012). The validity of teacher assessed Independent Research Reports contributing to Cambridge Pre-U GPR *Research Matters: A Cambridge Assessment Publication*, (14), 38-41.
Hughes, S. & Fisher-Hoch, H. (1997). Valid and invalid sources of difficulty in maths exam questions. *International Association of Educational Assessment.* South Africa.

Ipsos MORI. (2006). Teachers' views on GCSE coursework. Research study conducted for the QCA. Available: http://image.guardian.co.uk/sys-files/Education/documents/2006/10/06/teachersviewscoursework.pdf [Accessed 12th September 2012].

Ipsos MORI. (2011). Evaluation of the Introduction of Controlled Assessment. Report on qualitative and quantitative research. Ofqual/11/5049. Available: http://www.ofqual.gov.uk/files/11-10-05-Evaluation-of-the-Introduction-of-Controlled-Assessment.pdf [Accessed 12th September 2012].

Isaacs, T. (2010). Educational assessment in England. *Assessment in Education: Principles, Policy & Practice*, 17(3), 315-334.

Johnson, M. (2008). Exploring assessor consistency in a Health and Social Care qualification using a sociocultural perspective. *Journal of Vocational Education and Training*, 60(2), 173-187.

Kane, M. T. (2009). Validating the interpretations and uses of test scores. In: Lissitz, R. W. (ed.) *The Concept of Validity: Revisions, New Directions, and Applications.* USA: Information Age Publishing.

Messick, S. (1975). The standard problem: meaning and values in measurement and evaluation. *American Psychologist*, 30(10), 955-966.
Messick, S. (1989). Validity. *In:* Linn, R. (ed.) *Educational Measurement*. New York: Macmillan.

Messick, S. (1995). Validity of Psychological Assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749.

Morgan, C. (1996). The teacher as examiner: the case of mathematics coursework. *Assessment in Education: Principles, Policy & Practice*, 3(3), 353-375.

ISSN: 1758-2199

Morgan, C. & Watson, A. (2002). The interpretive nature of teachers' assessment of students' mathematics: issues for equity. *Journal for Research in Mathematics Education*, 33(2), 78-110.

Nadas, R., Suto, W. M. I. & Grayson, R. (2012). "Analyse this": How do teachers with differing subject specialisms interpret common assessment vocabulary? *British Educational Research Association,* University of Manchester, UK.

Oates, T. (2009). The Cambridge Approach Principles for designing, administering and evaluating assessment. Available: http://www.cambridgeassessment.org.uk/ca/digitalAssets/188934_cambridge_approach.pdf [Accessed 12th September 2012].

Pollitt, A. & Ahmed, A. (1999). A new model of the question answering process. *International Association for Educational Assessment*. Bled, Slovenia.

Pollitt, A., Entwistle, N., Hutchinson, C. & de Luca, C. (1985). *What Makes Exam Questions Difficult?*, Edinburgh, Scottish Academic Press.

QCA (2005). A review of GCE and GCSE coursework arrangements. London: Qualifications and Curriculum Authority.

Quinlan, T., Higgins, D. & Wolff, S. (2009). Evaluating the construct-coverage of the e-rater® Scoring Engine. *ETS RR-09-01*. Princeton, New Jersey: Educational Testing Service.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18, 119-144.

Shaw, S. & Suto, I. (2010). A tricky task for teachers: assessing pre-university students' research reports. *36th Annual Conference of the International Association for Educational Assessment*. Bangkok, Thailand.

Shaw, S. D., Crisp, V. & Johnson, N. (2011). A framework for evidencing assessment validity in large-scale, high-stakes international examinations. *Assessment in Education: Principles, Policy & Practice*, 19(2), 159-176.
Shay, S. (2006). The assessment of complex tasks: a double reading. *Studies in Higher Education*, 30(6), 663-679.

Shepard, L. A. (1993). Evaluating Test Validity. *In*: Darling-Hammon, L. (ed.) *Review of Research in Education*. Washington, DC: AERA.
Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher*, 36(8), 477-481.

ISSN: 1758-2199

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: history repeats itself again. *In:* Lissitz, R. W. (ed.) *The Concept of Validity: Revisions, New Directions and Applications*. USA: Information Age Publishing.

Stobart, G. (2009). Determining validity in national curriculum assessments. *Educational Research*, 51(2), 161-179.

Suto, I. & Shaw, S. D. (2010). A tricky task for teachers: assessing pre-university students' research reports. *Research Matters: A Cambridge Assessment Publication*, (10), 10-16.

Threlfall, J., Nelson, N. & Walker, A. (2007). Report to QCA on an investigation of the construct relevance of sources of difficulty in the Key Stage 3 ICT tests. London.

Tomlinson, M. (2004). *14-19 Curriculum and Qualifications Reform: Final Report of the Working Group on 14-19 Reform*., Annesley, Nottinghamshire, DfES Publications.

Vaughan, C. (1991). Holistic assessment: what goes on in the rater's mind? *In:* Lyons, L. H. (ed.) *Assessing second language writing in academic contexts* Norwood, NJ: Ablex Publishing Corporation.

Xi, X. (2008). What and How much Evidence Do We Need? Critical considerations in validating an automated scoring system. *In:* Chapelle, C. A., Chung, Y. R. & Xu, J. (eds.) *Towards adaptive CALL: Natural language processing for diagnostic language assessment.* Ames, IA: Iowa State University.

ISSN: 1758-2199