Example Response Sheet

For each row, circle the unit which is more demanding and indicate why the unit is more demanding using the appropriate domain and taxonomy information to explain your decision. If you struggled, include a question mark to indicate that you found the judgement difficult.

Domain	Unit		Why was the more demanding unit more demanding?
Affective	NVQ1	GCSE1	
	NVQ1	NVQ2	
	NVQ1	GCSE2	
	GCSE1	NVQ2	
	GCSE1	GCSE2	
	NVQ2	GCSE2	

The validity of teacher assessed Independent Research Reports contributing to Cambridge Pre-U Global Perspectives and Research

Jackie Greatorex Research Division **and Stuart Shaw** Cambridge International Examinations

Background

The Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999, p.9) frame test validity in terms of "the concept or characteristic that a test is designed to measure". That is, the Standards reflect a construct-centred approach to test validity. This perspective draws on the view that the theoretical, underlying construct such as mathematical aptitude, represented by an observable test score is the foundation for evaluating a test. Thus "all test scores are viewed as measures of some construct" (AERA, APA, NCME, 1999, p.174). The claim of validity is that the test adequately reflects the constructs and can be used as basis for the inference of attainment or aptitude depending on the test purpose.

It is important, therefore, to establish that tests elicit performances that reflect intended constructs and that test developers and providers have recourse to a reasonably well-informed and coherent theoretical model underpinning the construct(s) of interest if they are to operationalise aspects of the construct(s) for practical assessment purposes. In reality, however, "Tests are imperfect measures of constructs because they either leave out something that should be included... or else include something that should be left out, or both." (Messick, 1989, p.34). If the construct(s) is not well defined and test tasks are inappropriate, then it will be difficult to support claims an awarding body wishes to make about usefulness of its assessments, including claims that tests do not suffer from construct under-representation and construct irrelevance (CI).

The focus of this research is construct irrelevance. Its working definition for this study is that CI occurs when irrelevant constructs

systematically influence marks. A claim about construct relevance or irrelevance is an evidence-based judgement about the extent to which marks are interpreted as the skills and knowledge the assessment is intended to measure. The evidence includes the way in which the mark scheme is interpreted and applied.

Certain assessor behaviours are sources of construct irrelevance if they result in irrelevant constructs systematically influencing marks. It is important to stress that it is not the behaviours that necessarily result in CI – a behaviour is only a source of CI if marks are systematically influenced.

A review of literature relating to the assessment of school/college coursework or projects was undertaken to identify assessor behaviours that potentially affect assessment judgement. Research showed several behaviours, that is, assessors:

- compared a candidate's performance with another candidate's performance (Morgan, 1996; Crisp, 2010)
- expressed feelings towards a candidate (Vaughan, 1991; Crisp, 2010)
- laughed or noted amusement at a candidate or their performance (Vaughan, 1991; Crisp, 2010)
- predicted the quality of a candidate's future performance (Barritt *et al*., 1986; Crisp 2010)
- expressed a view on their own assessment practice (Crisp, 2010)
- commented on a candidate's characteristic such as skill/ability/gender (Barritt *et al*., 1986; Vaughan, 1991; Crisp 2010)
- used surface features of a candidate's work in judgements (Morgan and Watson, 2002)
- estimated a candidate's effort invested in the work (Crisp, 2010).

The Cambridge Pre-U Independent Research Report

The Cambridge Pre-U is an international post-16 qualification designed to prepare candidates to succeed at their university studies by fostering independent and self-directed styles of learning. It is administered by Cambridge International Examinations (CIE), and is available in a variety of subjects. The subjects include the Cambridge Pre-U Global Perspectives and Research (GPR). GPR comprises two components, Global Perspectives (GP) and the Independent Research Report (IRR). They are designed to be taught as two successive one-year courses with IRR building on GP.

The focus of this article is the IRR. The syllabus, not the present article, is the authoritative reference document (Cambridge International Examinations, 2008).

The IRR is designed to help candidates develop the ability to learn critically and independently. The candidate chooses a topic, develops a title in the form of a research question and undertakes research to write an essay or report which is 4,500 to 5,000 words long. The report should have an introduction identifying and exploring terms and issues, as well as stating the scope of the research and why the research is worth undertaking. The report should also have a conclusion. The report should be readable to the candidates' peers and the candidate should be able to explain it to a non-specialist in the subject area. The work should be independent and the candidate should engage intellectually with the sources of evidence (e.g. books, articles, the internet).

The candidate and teacher meet during the project to discuss the question and research. Teachers are active in determining the subject and scale of the report; they might conduct seminars/ workshops to discuss subject specific issues and approaches, later seminars might be used to share ideas. The teacher also assists candidates in identifying and locating sources of evidence, understanding and developing appropriate research methods and organisational skills. Once the candidate has chosen a research question there is minimum intervention by the teacher. There should be on-going opportunities in single or group tutorials to discuss progress. The role of the teacher is analogous to a higher education lecturer supervising an undergraduate dissertation.

IRR is assessed by the candidate's teacher. The assessment includes the report, a short (five to ten minute) terminal interview (viva) to authenticate the work is the candidate's, and the teacher's observations, experience and records of the candidate's progress in developing and producing the IRR.

Samples of marking are centre moderated by an internal moderator (IM) and externally moderated by CIE. External moderation checks the marking of the report using the mark scheme. (The IRR mark scheme is in the syllabus). It should be noted that part of the mark scheme is for assessing and internally moderating each candidate's Knowledge and understanding of the research process and Communication as evidenced by teacher's observations, experiences and records (AO1 and AO4a), see Table 1. That is, AO1 and AO4a are not externally moderated. This situation arises as external moderators do not have the teacher's and IM's experience of the candidate to judge AO1 and AO4a.

Teachers and IMs can make interim annotations and summary comments (hereafter annotations and comments) when marking the report. Outside of the IRR context markers make annotations and comments to explain decisions to others and support judgements during marking (Crisp and Johnson, 2007; Fowles, 2008). Therefore the IMs and EMs for IRR might find the annotations and comments useful. However, there is no requirement to make the annotations and comments in IRR marking.

Table 1: Assessment Objectives AO1 and AO4a

Assessment Objective	*Task*	*Clarification*
AO1 Knowledge and understanding of the research process	Design, plan, manage and conduct own research project using techniques and methods appropriate to the subject discipline	Knowledge of research methods and conventions Applies subject-specific knowledge to refine issue for investigation, identify question and conduct research. Own independent research using techniques and methods appropriate to the subject discipline i.e. literature search, relevant statistical/data handling and modelling techniques
AO4a Communication	Communicate clearly in negotiating and conducting the research project	Explanation and presentation of research methods, findings and conclusions

Note that AO4 was divided into AO4a and AO4b for the purposes of the research.

This study, which formed part of a wider programme of research into the IRR (Suto and Shaw, 2010; Shaw and Suto, 2010), investigated whether CI occurred when AO1 and AO4a were used to mark and internally moderate. It was conducted post hoc. A list of behaviours was taken from previous research, outside the Cambridge Pre-U context. The comments

made by the teacher and the IM on reports were searched for evidence of the behaviours. If several behaviours occurred, their relationship to marks was investigated. A systematic relationship between behaviours and marks would be indicative of CI.

Method

Data

92 candidates entered the IRR unit. The teacher and IM recorded comments on reports as appropriate. All available AO1 and AO4 comments were collected. There was a total of 150 comments: 67 comments about AO1 (60 from a teacher and seven from an IM) and 83 comments about AO4 (62 from a teacher and 21 from an IM). These comments were from a total of 70 candidates' reports from eight centres.

Qualitative coding

Two researchers developed a qualitative coding system to analyse the comments. The coding system was developed from the behaviours noted in the literature (Table 2).

Table 2: Coding categories and associated behaviours found in the literature

Behaviour noted in literature as a potential source of CI The assessor……	*Category description The teacher/IM……*
Compared a candidate's performance with another candidate's performance (Morgan, 1996; Crisp, 2010)	Compared a candidate's performance with another candidate's performance
Expressed feelings towards a candidate e.g. hostility (Vaughan, 1991; Crisp, 2010)	Expressed feelings towards a candidate
Laughed or noted amusement at a candidate or their performance (Vaughan, 1991; Crisp, 2010)	Expressed amusement at a candidate's performance/a candidate
Predicted the quality of a candidate's future performance (Barritt *et al*., 1986; Crisp 2010)	Predicted the quality of a candidate's future performance
Expressed a view on their own assessment practice (Crisp, 2010)	Expressed a view on their own summative assessment practice. NOT teaching/ formative assessment.
Commented on a candidate's characteristic such as skill/ability/ gender (Barritt *et al*., 1986; Vaughan, 1991; Crisp 2010)	Commented on a candidate demographic/ general ability
Used surface features of a candidate's work in judgements (Morgan and Watson, 2002)	Referred to a surface feature(s) of a candidate's work. NOT quality of written communication
Estimated a candidate's effort invested in the work (Crisp, 2010)	Estimated a candidate's effort invested in the work

Each comment was analysed for the presence/absence of each behaviour using the categories (Table 2). If part of a comment referred to a behaviour then the whole comment was categorised as referring to the behaviour. For example, the comment "Her desire to learn and complete this report was impressive from start to finish" arguably referred to the teacher "Expressing feelings towards a candidate". Therefore the entire comment, "CANDNAME was always very well informed, very well read, very focused and considered. Her desire to learn and complete this report was impressive from start to finish and she needed very little support from her supervisor", was allocated to the category.

An assessor and two researchers undertook the analysis. One researcher read the comments and coded each one according to the categories. The assessor blind coded the same comments. The assessor and researcher each made 1200 coding decisions. This provided two sets of coding for the same comments.

The coding by the researcher and the assessor was very similar; the agreement between them was very high. There were thirteen decisions out of 1200 when they disagreed. Instances of disagreement were passed over to the second researcher for adjudication. All comments were coded, none remained unresolved.

Results

In the final analysis a total of five out of the 150 comments referred to a behaviour in the coding scheme. These behaviours occurred whilst assessing five out of 70 candidates.

- Two comments were categorised as "Expressed feelings towards a candidate", they were:

  "CANDNAME was always very well informed, very well read, very focused and considered. Her desire to learn and complete this report was impressive from start to finish and she needed very little support from her supervisor." (Teacher's comment regarding AO1)

  "Communication highly effective both on paper and orally. Communicated in a mature and effective way with tutor. The candidate took considerable care to prepare for the meeting." (Teacher's comment regarding AO4)

- Three comments were categorised as "Referred to a surface feature(s) of candidate's work in judgements", they were:

  "Lack of intro and sustained argument – otherwise very good." (Teacher's comment regarding AO4)

  "Style is not academically formal in parts. Referencing not always clear or present. Clearly communicated well with tutor." (IM's comment regarding AO4)

  "Intro too long." (IM's comment regarding AO4)

Discussion

The research investigated whether CI occurred in the marking and internal moderation of AO1 and AO4a.

There were several limitations with the research. First, the list of behaviours was possibly inexhaustive. Second, the comments were a partial representation of each teacher's and IM's thoughts and deliberations. These limitations meant that some behaviours might be undetected. Third, the comments about AO4 did not differentiate between AO4a and AO4b. Therefore any behaviours related to AO4 cannot be attributed accurately to AO4a or AO4b. Despite these limitations the research evidence provides some useful and important findings.

Previous research, outside of the Pre-U context, identified several assessor behaviours which might be sources of CI if they systematically

influence marks. Two of these behaviours were noted in the comments. The teacher/IM "expressed feelings towards a candidate or a candidate's performance" and "referred to a surface feature(s) of a candidate's performance". The behaviours occurred in comments about five out of 70 candidates (i.e. five out of 150 comments). These findings have resonance with previous findings about coursework/ project work (Vaughan, 1991; Crisp, 2010; Morgan and Watson, 2002). The occurrence of these behaviours was **not** evidence of CI.

The lack of behaviours was a positive finding, particularly given that in several domains/professions erroneous information can influence judgement (Hackenbrack, 1992; Laming, 2004; Wistrich *et al*., 2005). Gaeth and Shanteau (1984) found that interactive training and practise reduced the influence of irrelevant information on the judgement of soil samples. Summers *et al*. (2004) found that formal education (rather than experience based learning) improved credit granting decisions. If experts in other domains can be trained to pay less attention to irrelevant information, then perhaps teachers and IMs can too. CIE runs standardisation meetings and provides other forms of centre support. This might well have contributed to the lack of behaviours found in the comments.

CI occurs only when such behaviours systematically influence marks. The lack of behaviours meant it was not possible to investigate a systematic relationship between the behaviours and marks. The result was that there was no evidence of CI.

There were different numbers of comments about AO1 (N=67) and AO4 (N=83). In other words some reports contained comments by the teacher/IM about AO1 but not AO4 and vice versa. This is not problematic as there was no requirement to make comments on reports, as noted earlier. However, it is interesting to consider why there were differences. The purposes of making annotations and comments on scripts might provide some insights. Crisp and Johnson (2007) and Fowles (2008) report two reasons for annotating and commenting whilst marking:

- Explaining decisions to others
- Supporting judgements and decision making during the process of marking

Perhaps there was a feeling that AO1 decisions were more self evident and needed fewer aide memoires than AO4 judgements.

Conclusion

This study found no evidence of CI in assessing the Knowledge and understanding of the research process (AO1) or Communication (AO4), and therefore no threats to validity were identified. This adds to the body of research supporting the teacher assessing the candidate's IRR performance (Suto and Shaw, 2010; Shaw and Suto, 2010), the validity of the teacher assessment and internal moderation. The findings suggest that AO1 and AO4 facilitate valid assessment. Furthermore, standardisation and other forms of centre support may be useful in guarding against CI.

References

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

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

Cambridge International Examinations (2008). *Cambridge Pre-U Syllabus- 2nd edition (UK)*. Cambridge International Level 3 Pre-U Certificate in Global Perspectives and Independent Research. For examination in 2010, 2011 and 2012. University of Cambridge Local Examinations Syndicate. http://www.cie.org.uk/qualifications/academic/uppersec/preu/subjects/subject/preusubject?assdef_id=1018

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

Crisp, V. & Johnson, M. (2007). The use of annotations in examination marking: opening a window into markers' minds. *British Educational Research Journal*, **33**, 6, 943–961.

Fowles, D. (2008). *Does marking images of essays on screen retain marker confidence and reliability?* A paper presented at the 34th Annual Conference of the International Association for Educational Assessment, 7–12 September 2008, Cambridge UK. http://www.iaea2008.cambridgeassessment.org.uk/ca/digitalAssets/135333_Microsoft_Word_-_Fowles.pdf

Gaeth, G. J., & Shanteau, J. (1984). Reducing the influence of irrelevant information on experienced decision makers. *Organizational Behavior and Human Performance*, **33**, 2, 263–282.

Hackenbrack, K. (1992) Implications of seemingly irrelevant evidence in audit judgement. *Journal of Accounting Research*, **30**, 1, 126–136.

Laming, D. (2004). *Human judgement: the eye of the beholder*. Thomson Learning: London.

Messick, S. (1989). Validity. In: R. Linn (Ed.), *Educational Measurement*. 13–103. New York: Macmillan.

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

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.

Shaw, S. & Suto, I. (2010). *A tricky task for teachers: assessing pre-university students' research reports*. A paper presented at the 36th Annual Conference of the International Association for Educational Assessment, 22–27 August 2010, Bangkok, Thailand. http://www.iaea2010.com/fullpaper/223.pdf

Summers, B. Williamson, T. & Read, D. (2004). Does method of acquisition affect the quality of expert judgment? A comparison of education with on-the-job learning. *Journal of Occupational and Organizational Psychology*, **77**, 2, 237–258.

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

Vaughan, C. (1991). *Holistic assessment: what goes on in the rater's mind?* In: L.H. Lyons (Ed.), *Assessing second language writing in academic contexts*. 111–125. Norwood, NJ: Ablex Publishing Corporation.

Wistrich, A. J., Guthrie, C. & Rachlinski, J. J., (2005). *Can judges ignore inadmissible information? The dfficulty of deliberately disregarding*. Cornell Law Faculty Publications. Paper 20. Available at http://scholarship.law.cornell.edu/lsrp_papers/20 accessed 22 March 2012