reasoning, or logic is that we ought to value thinking as an end in itself. We should value thinking, value our reason and rationality, as an excellence in itself; not as something that is simply the by-product of a particular academic discipline. On it depends our own autonomy. Yes, it does underlie specialist subjects – so it will (and does) enhance what is done in each of those. But more importantly, it underlies what it means to be human.

A tricky task for teachers: assessing pre-university students' research reports

Irenka Suto Research Division **and Stuart Shaw** CIE Research

Introduction

In the UK and internationally, many students preparing for university are given the challenge of conducting independent research and writing up a report of around 4000 or 5000 words. Such research activities provide students with opportunities to investigate a specialist area of study in greater depth, to cross boundaries with an inter-disciplinary enquiry, or to explore a novel non-school subject such as archaeology, cosmology or anthropology. We theorise that, as is the case in higher education (Brown *et al*. 1997), independent research encourages intellectual curiosity whilst enabling students to develop skills in practical and analytical research, higher order thinking, interpretation and time management. When applying to university, students can use their reports to demonstrate motivation for their intended course of study and to differentiate themselves from competing applicants.

In the wake of the recommendations of the Tomlinson Report (2004) on the shape of 16–19 qualifications in England, The Sixth Form College, Farnborough, developed a systematic approach to encouraging its students to conduct independent research. Since 2006, students have been carrying out extended projects during their holidays or alongside their other courses, generating formally-structured reports. The reports are assessed formatively through detailed written comments to the students by their teachers, rather than assessed summatively by issuing a mark. This has generated a considerable body of student evidence within the college.

At other schools, students conduct projects which constitute or contribute to a formal qualification, and which are therefore assessed summatively. For some of these qualifications, the students' research reports are assessed by their own teachers. The teachers' marks are then moderated by professional examiners who are employed by the examination board administering the qualification. The Cambridge Pre-U Independent Research Report, administered by Cambridge International Education, utilises this assessment approach, as do the extended projects administered by the AQA, OCR, and Edexcel examination boards. Extended projects can be used to obtain a stand-alone qualification. Alternatively they can contribute to a 14–19 Diploma in England or the Welsh Baccalaureate qualification in Wales. For other qualifications, such as the International Baccalaureate, students' research is marked exclusively by external examiners.

The assessment of research reports poses several challenges, including those which arise when assessment schemes are designed to reward generic research skills rather than particular subject knowledge. Assessors may lack detailed understanding or marking experience of the research topics explored by some students. However, it is unclear whether subject knowledge facilitates or hinders marking. For example, familiarity with particular terminology or technical language may aid interpretation of what the student has written. Alternatively it may obscure the assessor's perception of generic skills, especially if they have been mis-applied by the student.

In this study, we explored the feasibility of applying a single mark scheme to research reports covering diverse topics in order to reward generic research skills. Our aim was to investigate the reliability with which teachers can mark diverse research reports, using four different generic assessment objectives. We also investigated teachers' views in applying generic mark schemes, particularly when marking reports on unfamiliar topics.

The Cambridge Pre-U Independent Research Report (IRR)

The study was conducted as part of a wider on-going research programme supporting the Cambridge Pre-U, a new type of qualification for 16–19-year-olds which is designed to equip students with the skills required to make a success of their university studies. The first cohort of Cambridge Pre-U students will be completing their courses in the summer of 2010. Typical Cambridge Pre-U students 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 but also constitutes a stand-alone qualification with a UCAS tariff equivalent to an A level. It comprises two components: the Global Perspectives course (GP), and the Independent Research Report (IRR) which may be up to 5000 words long. The GP and IRR have been designed to provide students with coherence, depth and breadth, through encouraging focused personal exploration and increased depth of study. They expand creative, critical and responsible awareness through the tackling of different perspectives on global issues. Assessment of the IRR focuses on the student's abilities in a range of areas. These include: designing, planning and managing a research project, collecting and analysing information,

evaluating and making reasoned judgements, communicating findings and conclusions, and uniquely, intellectual challenge. The present study explores the practical application of four different generic assessment objectives which comprise a substantial proportion of the mark scheme that will be used to mark the IRR this summer.

Participants

Fifteen teachers (10 men and 5 women) participated as markers in the study. They were recruited by e-mail from nine different schools in England whose 16–19 year-old students were either currently working on independent projects or planning to do so in the near future. The teachers had a wide range of subject backgrounds and teaching and examining experiences.

The teachers' experimental marking was led by a highly experienced examiner: the Chief Examiner (CE) for Cambridge Pre-U's GPR course, who also undertook this role in the study.

Project reports

The study was conducted prior to the completion of any Cambridge Pre-U IRRs by Cambridge Pre-U students. We therefore explored the marking of project reports produced by students of The Sixth Form College, Farnborough, UK. Like IRRs, the projects could be on any topic of interest to students, the reports had an approximate word limit of 5000 words. However, as the projects did not contribute to any qualification, the students had not written the reports with any particular assessment objectives or marking criteria in mind.

The college provided the researchers with copies of 346 project reports (68 from 2006, 135 from 2007, and 143 from 2008). At a two-day meeting, the researchers and CE jointly reviewed the reports and selected a sample of 20, stratified by subject area. From these 20 reports, a sub-sample of 5 was selected for use by participating teachers as a practice sample. Full details of the report selection process are given in the appendix.

The CE determined a fixed marking order for the 5 reports in the practice sample, which were numbered accordingly. The remaining 15 reports comprised the main sub-sample. The researchers determined a random marking order for these reports and numbered them accordingly. The report titles are shown in Table 1.

Mark scheme

An experimental version of a mark scheme was used in the study which was derived from that for the Cambridge Pre-U IRR. The original IRR mark scheme is divided into five Assessment Objectives (AOs, see Table 2), enabling assessment of each of the five AOs at three different levels. Since for AO1, students are required to "design, plan, manage and conduct own research project using techniques and methods appropriate to the subject discipline", AO1 can only be assessed in the context of the classroom, by students' own teachers. As the study's teachers were to mark the work of students they had not taught, AO1 was omitted in the experimental mark scheme. Similarly, part of AO4 relates to a student's negotiation with his/her tutor; as it could not be used in this study, it was removed from the experimental mark scheme.

Table 1: Titles of project reports used in the study

Sub-sample	*Report number*	*Project report title*	*Broad subject area*
Practice	01	Can we trust Quantum Theory over Electromagnetic Wave Theory of Light?	Physics
	02	Would the British economy have been as successful without the transatlantic slave trade?	Economics
	03	Is prison the best sentence for paedophiles, or do alternatives offer a safer and more effective rehabilitation option?	Criminology
	04	Addiction – nature or nurture?	Psychology
	05	Polya's heuristics: are they applicable in a broader context?	Mathematics
Main	06	How effectively has Ghana dealt with the problem of malaria?	Geography
	07	An exploration into the role of metaphor in economics	English
	08	Is prescribed medication the most effective way to treat Attention Deficit Hyperactivity Disorder?	Biology
	09	Does the French language need protecting, and if so is enough being done to protect it?	French
	10	Is it right to chemically alter the behaviour of children through the use of drugs such as Ritalin?	Biomedical ethics
	11	Has Pina Bausch revolutionised ballet with her controversial 'Tanztheater'?	Drama
	12	Hydrogen fuel: can hydrogen replace gasoline?	Chemistry
	13	Should the UK join the Euro?	Politics
	14	Could an artificial intelligence be an ideal ruler?	Philosophy
	15	Can a murderer's behaviour be reduced down to biological or environmental factors, or is it a combination of both?	Psychology
	16	Is communism viable today?	Politics
	17	Is punk rock art?	Art
	18	Should permission be given to remove the treatment of patients in a persistent vegetative state?	Biomedical ethics
	19	What philosophical problems arise with Chomsky's account of language acquisition?	Linguistics
	20	To what extent does music have a beneficial effect on brain activity?	Music

Table 2: Assessment objectives and marks in original mark scheme

Assessment Objective	*Domain*
AO1	Knowledge and understanding of the research process
AO2	Analysis
AO3	Evaluation
AO4	Communication
AO5	Intellectual challenge

Procedure

The experimental procedure comprised the following stages:

1. The Chief Examiner (CE) marked all 20 reports, thereby generating a 'correct' mark for each one.
2. Each teacher was posted the sample of 20 numbered reports, together with the mark scheme, practical instructions about the study from the researchers, and detailed written guidance on marking from the CE. A marking grid was also provided, to be used to record marks and notes.
3. Each teacher began by familiarising him/herself with the mark scheme and reading the CE's guidance on marking.
4. Each teacher marked the practice sub-sample (N = 5) in numerical order, recording his/her level followed by his/her mark and notes in the marking grid. Teachers were welcome to annotate the reports.
5. Each teacher contacted the CE, who provided personalised telephone feedback on his/her marking of the practice sample. Teachers were asked to record the CE's marks and feedback in their marking grids. The CE also kept records of the teachers' marks and the feedback given.
6. After receiving telephone feedback, each teacher marked the main sample (N = 15) in numerical order. The teachers were asked to try to apply the CE's advice wherever possible. For each report, they recorded their marks and notes for each assessment objective in the marking grid. Again, the teachers could annotate the reports if they wished.
7. After completing the marking, each teacher filled in a questionnaire about his/her marking experiences.
8. All documents were returned to the researchers.

Analysis and findings

All 15 teachers marked all 20 reports in the study. However, one teacher had to withdraw from the study for personal reasons prior to completing the post-marking questionnaire. Analyses were conducted on the marking of the main sub-sample and the questionnaire data using SPSS Version 15.01 and FACETS Version 3.6 software.

Correlation of marks

Indices of inter-rater reliability among all participants (i.e. the 15 teachers and the CE) were calculated for each of the four Assessment Objectives (AO2–5) and for the total score using a procedure described by Hatch and Lazaraton (1991, p.533). This entailed generating a Pearson correlation matrix for all participants for each AO. A Fisher Z transformation was then applied to the correlations, to transform the correlations to a Normal distribution and to correct the distortion inherent in using the Pearson for ordinal data. The mean correlation among participants could then be calculated. Subsequently, the derived mean of the transformed correlation coefficients, r_{ab} was substituted into the formula:

$$r_{tt} = \frac{n.r_{ab}}{1+(n-1)r_{ab}}$$

where r_{tt} stands for the reliability of all the participants' ratings, n is the number of participants, and r_{ab} is the average correlation among participants. Finally, r_{tt} was transformed back to a Pearson's correlation coefficient.

Table 3: Inter-rater marking reliabilities (among all participants)

	Number of marks available	*Pearson's correlation coefficient*
AO2	18	0.71
AO3	18	0.72
AO4	9	0.71
AO5	6	0.73
Total score	**51**	**0.72**

Table 3 presents the mean correlations for each AO and for the total score.

These reliability figures compare favourably with those estimated and reported elsewhere. For example, Shaw (2008) quotes inter-rater reliability indices of 0.78 using the same statistical approach. In another, similar study investigating marking reliability of essay questions from the higher tier of GCSE English Literature, Johnson, Nádas and Bell (2009) also report reliabilities of a comparable magnitude. However, these studies both focus on medium length constructed responses which are considerably shorter than the 5000-word reports used in the present study. The focus of a study by Laming (1990) offers a closer comparison. Laming's investigation was designed to estimate reliability between pairs of examiners marking a university examination comprising a number of extended essay-type answers. Laming found that the correlation between the marks independently awarded by pairs of examiners varied between 0.13 and 0.72. Given the participants' lack of familiarity with the present study's experimental mark scheme, the reliability figures calculated here are encouraging.

These findings were corroborated by a statistical check employing multi-faceted Rasch analysis. In the context of inter-rater reliability, FACETS models participants as 'independent experts'. Although FACETS does not estimate inter-rater reliability directly, it routinely generates observed and expected agreement percentages. Adapting Cohen's Kappa agreement statistic enables the estimation of a Rasch-based Kappa coefficient. Under Rasch-model conditions ideally this should be close to 0, indicating that inter-rater reliability is within the acceptable range.

The Rasch-Cohen's Kappa is calculated as:

$$\frac{(\text{Observed agreement \%} - \text{Expected agreement \%})}{(100 - \text{Expected agreement \%})}$$

Values of Rasch-Cohen's Kappa for each AO are presented in Table 4.

Table 4: Values of Rasch-Cohen's Kappa for AOs

Assessment Objective	*Rasch-Cohen's Kappa*
AO2	0.0088
AO3	0.0212
AO4	0.0038
AO5	0.0160

These values are close enough to 0 to support the previous findings of high reliability for report marking.

In order to explore participant agreement further, FACETS was used to provide two measures of 'fit' (or consistency): the 'infit' and the 'outfit'

values.[1] There are different views on what fit index is actually acceptable. McNamara (1996) suggests that the usual limits of acceptability are the mean ± 0.3 (so anything between 0.7 and 1.3 is acceptable). According to Lunz and Wright (1997, p.83) "Because the interpretation of fit is situationally dependent, there are no fixed levels for fit statistic acceptance or rejection." They go on to use a level of ± 0.5 in their studies. Wright and Linacre (1994, p.370) suggest figures ranging from 0.4 for 1.7 depending on the type of assessment under investigation: fit statistics of 1.7 or greater indicate too much unpredictability in raters' marks, while fit statistics of 0.4 or less indicate overfit or not enough variability in raters' marks. The infit and outfit values for the CE and 15 teachers were calculated for each AO. Overall, given the above guidance on levels of fit, they indicated a generally well-fitting Rasch model.

When considered together with the descriptive statistics and estimations of inter-rater reliability, the Rasch findings reveal a good degree of agreement among participants on each of the four AOs.

Relative marking severity and variation

For each report, the CE's marks were deemed to be correct and therefore the 'gold standard'; they were used as the comparators against which all teachers' marks were compared. This analysis explored marking agreement with a consideration of two descriptive statistics:

- *marking mean* – a measure of relative severity of the marking.
- *standard deviation* – a measure of the range of marks used. The larger the standard deviation, the wider the range of marks awarded.

Table 5 summarises the mean total marks given by each participant to the 15 reports. On average, the CE's total marks are lower than those awarded by the teachers and cover a narrower range. ANOVA revealed a significant difference among the participants ($F = 2.36$, d.f. = 15, 224, $p < 0.05$); however, deeper investigation with post-hoc tests (Bonferroni and Tukey) indicated that only one teacher (G) marked significantly more severely than the others.

An analysis of the marks awarded on individual assessment objectives was also conducted. Both AO2 (Analysis) and AO3 (Evaluation) employ a mark range of 1–18 marks across three levels. The mean marks in Table 6 and Table 7 indicate the relative severities of the 15 teachers and CE on these two AOs.

Table 5: Descriptive statistics for the total marks given by participants

Teacher	*Main subject(s) taught*	*Mean mark*	*Standard deviation*
CE	History	26.93	9.05
A	Critical thinking	31.60	9.98
B	History, politics, business studies	25.07	11.95
C	Law, politics, psychology	28.20	9.44
D	History	28.67	10.55
E	Religious studies, philosophy	33.27	9.07
F	Philosophy, ethics, religious studies	31.07	8.96
G	Physics, astronomy	22.93	8.36
H	English, media studies	31.53	9.58
I	English	29.07	8.48
J	Maths	34.53	10.72
K	Politics, history, critical thinking	25.73	9.84
L	Biology, chemistry	23.27	10.77
M	Theory of knowledge, classical civilisation	33.40	10.24
N	English, critical thinking	35.67	7.58
O	Chemistry	28.60	11.97

1 The infit is the weighted mean-squared residual (the difference between actual marks and marks predicted by the Rasch model) which is sensitive to unexpected responses near the point where decisions are being made, while the outfit is the unweighted mean-squared residual and is sensitive to extreme scores. For ease of interpretation, the two sets of fit statistics are expressed either as a mean square fit statistic or as a standardised fit statistic, usually a *z* or *t* distribution.

Table 6: AO2 Descriptive statistics

Teacher	*Mean mark*	*Standard deviation*	*N reports marked*
CE	8.20	3.28	15
A	10.33	4.06	15
B	8.53	4.45	15
C	9.60	3.64	15
D	10.33	3.83	15
E	11.40	3.58	15
F	10.80	3.19	15
G	6.13	2.61	15
H	10.13	3.66	15
I	8.80	3.28	15
J	11.47	3.81	15
K	7.73	3.79	15
L	7.60	3.98	15
M	10.93	3.90	15
N	12.67	2.82	15
O	9.07	4.62	15

Table 7: AO3 Descriptive statistics

Teacher	*Mean mark*	*Standard deviation*	*N reports marked*
CE	8.93	3.99	15
A	10.73	4.64	15
B	8.20	4.48	15
C	10.00	3.44	15
D	8.47	4.73	15
E	11.53	3.04	15
F	10.53	3.36	15
G	8.40	3.44	15
H	10.80	3.99	15
I	10.40	3.89	15
J	11.80	3.84	15
K	9.53	4.22	15
L	8.27	4.40	15
M	11.40	3.72	15
N	11.93	3.17	15
O	9.60	4.78	15

For AO2, the mean marks ranged from 6.13 to 12.67. ANOVA revealed significant differences among the participants ($F = 3.24$, d.f. = 15, 224, $p < 0.05$); post-hoc tests indicated that Teachers G, K and L marked significantly differently from the others. The table shows a spread in standard deviation of nearly 2 marks when assessing AO2.

Whilst there were differences in severity among teachers in the marks awarded for AO3, these were less marked than for AO2 and not statistically significant ($F = 1.61$, d.f. = 15, 224, $p > .05$), that is, the participants generally behaved as a homogeneous group. Although AO3 and AO2 are equally weighted, the tables reveal a greater spread of marks for AO3, suggesting that AO3 is discriminating among reports more effectively.

In general, the CE tended to mark more harshly on both AO2 and AO3 than the teachers do, although this tendency is less pronounced on AO3 and over a slightly narrower range on AO2.

AO4 (Communication) is assessed against a 9 mark scale. As Table 8 shows, the trend towards CE severity (apparent for AO2 and AO3) is reversed in the case of AO4 where teachers tended to be slightly more severe than the CE.

AO5 is assessed against a 1– 6 mark scale, which is the shortest scale. Evidence from the marks (Table 9) suggests that, on average, the CE marked more harshly on AO5, and over a slightly wider range, than the

Table 8: AO4 Descriptive statistics

Teacher	*Mean mark*	*Standard deviation*	*N reports marked*
CE	6.73	1.71	15
A	6.47	1.73	15
B	5.40	2.10	15
C	5.80	1.42	15
D	6.47	1.41	15
E	6.33	1.72	15
F	5.73	1.67	15
G	5.53	1.85	15
H	6.60	1.24	15
I	6.60	1.76	15
J	6.93	1.94	15
K	5.73	1.49	15
L	5.07	2.12	15
M	6.87	1.68	15
N	7.00	1.31	15
O	5.87	2.23	15

Table 9: AO5 Descriptive statistics

Teacher	*Mean mark*	*Standard deviation*	*N reports marked*
CE	3.07	1.39	15
A	4.07	1.28	15
B	3.00	1.51	15
C	2.80	1.26	15
D	3.40	1.72	15
E	4.00	1.25	15
F	4.00	1.13	15
G	2.87	1.41	15
H	4.13	1.13	15
I	3.27	1.03	15
J	4.33	1.50	15
K	2.80	0.94	15
L	2.47	1.13	15
M	4.20	1.37	15
N	4.07	0.88	15
O	3.80	1.86	15

teachers. As with AO2, ANOVA revealed significant differences among the participants ($F = 3.28$, d.f. = 15, 224, $p < .05$); post-hoc tests indicated that Teachers J, L and M marked significantly differently from others.

The scatter diagram in Figure 1 shows the relationship between the mean of the teachers' total marks and the CE's (gold standard) total marks. If the two marking approaches were to yield identical marks, then the points on a scatter diagram would all lie on a *line of identity*, shown with a dotted line in Figure 1. It can be seen that ten points lie above the identity line, indicating frequent marking leniency relative to the CE reports. Very few points lie below the identity line, indicating that marking severity relative to the CE was much rarer.

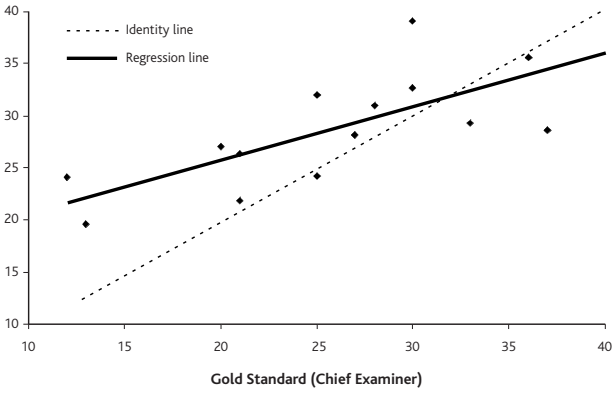

Figure 1: Comparison of CE's 'gold standard' total marks with mean teacher total marks

It can be seen that the regression line (bold line) is generally less tilted than the line of identity, showing that the teachers as a group tended to be less likely to use the extremes of the mark scheme than the CE. However, this could be interpreted as evidence of regression to the mean, as individually, the teachers used wider ranges of marks than the CE did.

Discussion

The above analyses indicate that marking reliability was good, though like almost all qualifications (Suto, Nádas and Bell, 2009), imperfect. Possible reasons and explanations for marking difficulty were identified by the participating teachers, which were recorded as written comments in their marking grids and questionnaire responses. Table 10 and Table 11 summarise the teachers' explanations for why some reports were harder and easier to mark than others.

The teachers' comments indicate that many of them found it easier to mark reports within their own subject areas, despite the generic nature of the Cambridge Pre-U IRR mark scheme. Subject knowledge appears to have facilitated some teachers' understanding of the language and terminology used. However, this experience was by no means universal, with one teacher commenting that clarity of thought was critical to marking ease, even with research reports on alien subject matter. Moreover, one teacher gave having 'too much subject knowledge' as a reason for finding some reports harder to mark than others. It may be that for this particular teacher, subject knowledge obscured his or her perception of generic skills. Other comments from the teachers point towards individual differences in perceptions of what affects marking difficulty: whilst one teacher felt that good performances were easier to mark, another teacher felt that poor performances were easier to mark.

The teachers' comments provide a useful window into the nature of research report marking. However, it is worth noting that perceived marking difficulty is not the converse of marking accuracy. A marking task may feel difficult without accuracy necessarily being compromised, since assessors may put greater effort into demanding marking situations, as found by Johnson, Nádas and Bell (2009). Similarly, marking confidence may not be a good indicator of actual marking accuracy, since genuine insight into the marking process may be lacking, as has been found to be the case for some GCSE examiners (Nádas and Suto, 2007).

To conclude, the levels of marking reliability found in this study are encouraging. This is especially so given the study's limitations, which include the unavailability of authentic Cambridge Pre-U independent research reports, the novelty of the mark scheme, and the inexperience of the teachers involved in this study, who had no prior training and no access to material exemplifying standards. Future challenges for researchers include exploring assessment objectives that can only be assessed in the context of the classroom, by students' own teachers. Not all research skills can be assessed via a written research report and it is important that skills such as knowledge and understanding of the research process (AO1 in the Cambridge Pre-U's IRR mark scheme) can also be rewarded consistently.

Table 10: Perceived reasons for difficulty of marking some reports

Perceived reasons for finding some reports harder to mark than others	*Illustrative quotes from teachers*
Main reasons • Technical language and terms; lack of background/specialist knowledge (N = 8) • Density of language (N = 4) **Other reasons** • Evaluating quality of sources of information • Intellectually challenging • Discerning structure/arguments • Lack of proper evaluation • Too much subject knowledge	*"There was a lot of technical language upon which the arguments and analysis were based. One needed to keep all of these new technical terms in mind whilst trying to assess how effectively the sources and perspectives had been dealt with. It felt a bit like spinning plates, with constant shuffling from one part of the project to another to check for meanings and consistency of their use."* *"The critical thinking and evaluative aspects were tricky to pick out of the density of the text."* *"Not only was this far from my 'home area', but the terminology was foreign."*

Table 11: Perceived reasons for ease of marking some reports

Perceived reasons for finding some reports easier to mark than others	*Illustrative quotes from teachers*
Main reasons • Within subject area (N = 7) – taught – studied – familiarity – academic specialism • Clear analysis of perspectives; clarity of thought/argument/terminology (N = 5) **Other reasons** • Easy to judge use of source material • Short • Poor performance • Good performance • Marking familiarity – increased during course of study	*"…on a topic I have in-depth knowledge of."* *"It was easiest for me to mark the report on Communism as that is closest to my own academic specialism."* *"The ones which were easiest to mark were the reports presented with clarity of thought, even though the subject matter was unfamiliar to me."* *"…in my comfort zone of an area plus it was clearly argued and debated with discussion of the main criteria reflected in the mark scheme e.g. the notion of flaw etc."* *"…because it was easy to read, relatively short and at a low level."*

Acknowledgments

We are very grateful to The Sixth Form College, Farnborough for the use of past students' project reports. We would also like to thank the teachers who participated in this study.

References

Brown, G., Bull, J. & Pendlebury, M. (1997). *Assessing student learning in higher education*. Routledge: London and New York.

Hatch, E. & Lazaraton, A. (1991). *The Research Manual: Design and Statistics for Applied Linguistics*. 533–535. Boston, Massachusetts: Heinle & Heinle.

Johnson, M., Nádas, R. & Bell, J. F. (2009). Marking essays on screen: An investigation into the reliability of marking extended subjective texts. *British Journal of Educational Technology*, published online.

Laming, D. (1990). The reliability of a certain university examination compared with the precision of absolute judgments. *Quarterly Journal of Experimental Psychology*, **42A**, 239–254.

Lunz, M.E. & Wright, B.D. (1997). Latent Trait Models for Performance Examinations. In: Jürgen Rost and Rolf Langeheine (Eds.) *Applications of Latent Trait and Latent Class Models in the Social Sciences*. http://www.ipn.uni-kiel.de/aktuell/buecher/rostbuch/ltlc.htm

McNamara, T.F. (1996). *Measuring Second Language Performance*. London: Longman.

Nádas, R. & Suto, I. (2007). An exploration of self-confidence and insight into marking accuracy among GCSE maths and physics markers. (A magabiztosság és a teljesítménybecslés pontosságának kutatása az angol GCSE vizsgák értékelo˝ inél') *Magyar Pedagogia*, **107**, 3, 169–184.

Suto, I., Nádas, R. & Bell, J.F. (2009). Who should mark what? A study of factors affecting marking accuracy in a biology examination. *Research Papers in Education*, published online.

Shaw, S.D. (2008). Essay Marking On-Screen: implications for assessment validity. *E–Learning*, **5**, 3, 256–274.

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

Wright, B. & Linacre, J. (1994). Reasonable Mean-square Fit Values. *Rasch Measurement Transactions*, **8**, 3, 370.

Appendix: Stages in the two-day project report selection meeting

1. The CE read through all 346 project report titles and categorised them as either *no* or *yes/maybe* according to whether they would be verified as Cambridge Pre-U Independent Research Report titles. The main criterion for rejection was that a title did not explicitly invite discussion. Only titles that seem to lead to discussion are suitable for the IRR. There were 118 *yes/maybe* reports in total.
2. The CE and researchers looked at the participating teachers' subject backgrounds and made a list of all subject specialisms. Any missing major subjects (e.g. geography, psychology) were added to the list. The list was then revised and refined to form broad 16 subject areas, into which the project reports could probably be grouped.
3. The CE and researchers grouped the *yes/maybe* reports into the 16 subject areas.
4. The initial subject classifications of each report were checked, subject area by subject area, in a group discussion. Some reports were moved to different subject areas at this point. The numbers of reports in each subject area were counted (N = 118 in total).
5. The report titles in each subject area were checked (again in a group discussion) and the CE discarded any reports with titles that he did not think he could ultimately verify. The numbers of reports in each subject area were counted (N = 94 in total).
6. In a discussion of how to select the 20 reports needed for the study, the CE proposed that the criteria for spotting top reports would be:
   - Incisive conclusions (AO3)
   - Alternative interpretations (AO3)
   - Uses a range of sources (AO2)
   - Critical vocabulary (AO4)

   It was agreed that the first 5 reports that teachers mark should flag up key issues that need to be addressed in the CE's feedback.
7. The CE and researchers read through the reports in the subject areas (each taking the subject areas that s/he knew most about) and selecting one or two possible reports for inclusion in the sample on the basis of them being (1) very strong, (2) very weak, or (3) interesting and likely to generate discussion. This generated a selection of 23 reports.
8. Three reports from the most over-represented subject areas (economics, history and geography) were excluded from this selection to leave 20 reports.
9. The CE suggested that the practice sub-sample of reports should help the participating teachers to understand the marking criteria by illustrating key aspects of the mark scheme. The CE identified the following selection requirements:

   Report 1: AO2 and AO3 at level 3
   Report 2: AO2 and AO3 at level
   Report 3: AO2 and AO3 at level 1 or 2
   Report 4: AO5 at level
   Report 5: AO2, AO3, AO4 and AO5 at level 2.
10. In a group effort, five reports were found which met the above requirements and also covered a good mix of subject areas. They were then ordered so that they would not be encountered in either ascending or descending order of quality, but in a mixed order of quality.

Details of the selection process are summarised in Table A1.

Table A1: Details of the report selection process

Subject area	*Reports placed in each subject area after initial verification of title by Chief Examiner as 'yes/maybe' (N = 118)*	*Reports placed in each subject area after final consideration of titles (N = 94)*	*Reports initially selected for full sample of 20 (N = 23)*	*Reports finally selected for full sample of 20 (N = 20)*	*Reports used in the IRR marking study (N = 20)*	
					Reports selected for the main sub-sample (N = 5)	*Reports selected for the practice sub-sample (N = 15)*
Art & architecture	2	2	1	1	0	1
Biology	11	6	1	1	0	1
Biomedical ethics	11	10	2	2	0	2
Chemistry	2	2	1	1	0	1
Economics	10	8	1	1	0	1
English & applied linguistics	7	5	2	1	0	1
French	4	3	1	1	0	1
Geography	5	5	2	1	0	1
History	6	6	2	1	1	0
Law	8	7	1	1	1	0
Maths & computing	4	4	1	1	1	0
Music, film & drama	7	5	2	2	0	2
Philosophy & religious studies	7	5	1	1	0	1
Physics & astronomy	7	4	1	1	1	0
Politics	9	7	2	2	0	2
Psychology & sociology	18	15	2	2	1	1

Towards an understanding of the impact of annotations on returned examination scripts

Martin Johnson Research Division **and Stuart Shaw** CIE Research

Introduction

For the past few years awarding bodies in England, Wales and Northern Ireland have been obliged to allow assessment centres and candidates to request to see their examination scripts once they have been marked. Guidelines established by the regulator of qualifications in England, the Office of the Qualifications and Examinations Regulator (Ofqual) in conjunction with the Welsh Assembly Government's Department for Children, Education, Lifelong Learning and Skills (DCELL) and the Northern Ireland Council for Curriculum, Examinations and Assessment (CCEA) outline the steps that qualification awarding bodies need to take to ensure that this accountability function is fulfilled.

According to these documents centres and individual assessment candidates have the right to access marked examination scripts under certain conditions which safeguard issues of candidate data confidentiality. There is little empirical study into practices around scripts returned to centres. It appears intuitive that script requests might be considered as a precursor to a results enquiry but what is less intuitive is whether any other uses are made of these returned scripts.

Returned scripts often include information from examiners about the performance being assessed. As well as the total score given for the performance, additional information is carried in the form of the annotations left on the script by the marking examiner. As far as we know there has been no research into how this information is used by