Discrepancy essays – natural phenomenon or problem to be solved?

David CONIAM

The Chinese University of Hong Kong

Abstract

This paper describes a study investigating discrepancies between raters on an English language Writing Test public examination in Hong Kong where paper-based marking (PBM) in public examinations is soon to be replaced completely by onscreen marking. The discovery and development of the phenomenon of discrepancy scripts arose from an analysis of data in a related study (Coniam, 2009) in which 30 raters rated, on paper, scripts they had at some time previously rated on screen. In that study, ratings on a number of scripts revealed that at least one rater had rated them more than 20% (5/24 points) more severely on paper than on screen, while at least one other rater had rated the same scripts more than 20% (5/24 points) more severely on screen than they had on paper. The current study was therefore constructed to investigate these 'discrepancy scripts' in an attempt to discover whether specific causes of such discrepancies could be identified. A set of 15 scripts were identified across the two marking mediums. These scripts had received grades diverging by 5/24 points or more. A further control set of scripts was also identified comprising 15 scripts which had received exactly the same grade from two different raters. 12 raters were used in a crossed design. Six rated the discrepancy scripts on screen while the other six rated them on paper. The two groups of raters then changed around, rating the other set of scripts in the other medium. While rating, they noted whether particular scripts were easy or difficult to rate. After the rating exercise, the raters took part in semi-structured interviews. From the analysis involving multi-faceted Rasch measurement, expectations that the discrepancy scripts would show greater misfit in the Rasch model than the same-grade scripts were not borne out. Likewise, raters' evaluations, based on two topics in different genres, showed no bias in the discrepancy scripts, nor did any features emerge which might allow a definition of discrepancy scripts to be developed. The paper concludes that some variation may be inevitable and may have to be accepted in any rating situation.

Melbourne Papers in Language Testing 2009 Volume 14.2 pp. 1-31. The Language Testing Research Centre, The University of Melbourne.

Introduction

In Hong Kong, paper-based marking (PBM) in public examinations is imminently to be phased out and replaced totally by onscreen marking (OSM) in 2012. To investigate, and validate, the adoption of OSM as the future sole method of marking, a series of studies is being undertaken to compare the two modes of rating – utilising one of the Year 11 (Secondary 5) English language public examinations, the 2007 Hong Kong Certificate of Education (HKCE) examination English language Writing Paper.

Onscreen marking has developed rapidly in the past decade or so with a number of studies comparing two different marking modes (Powers et al. 1997; Powers & Farnum, 1997; Powers et al., 1998). In the Powers et al. (1997) and the Powers and Farnum (1997) studies when experienced raters scored essays both on paper and on screen, no differences emerged between the average scores awarded in either medium and inter-rater correlations were comparable for both methods. Scores were not affected by the medium in which essays were presented to readers – on screen or on paper.

In the UK, Newton et al. (2001) and Whetton & Newton (2002) evaluated the online marking of Year 7 progress tests, employing both expert and non-expert raters and found, generally, that there was no difference in the overall ratings of either group. Sturman & Kispal (2003) also found that no consistent trends emerged in the differences in test scores between the two modes of rating.

Zhang et al. (2003: 21) found the agreement between OSM raters on essay questions was "at least as good as that for those who read in a traditional operational setting" and Adams (2005), discussing the analysis of the marking of individuals test items as against whole scripts, reported no significant differences between OSM and PBM.

In Powers et al. (1997), the OSM system was received "relatively positively" (p. 10) by most raters, despite the fact that some raters had misgivings about the system. In Zhang et al.'s (2003) study, raters' reactions were mixed: some who were less positive drew attention to: the lack of opportunities to discuss issues with other raters; the lack of printed commentaries on training essays andhaving to scroll down the screen in order to read some essays. Raters' reactions to most aspects of OSM were generally satisfactory, although a "significant minority" of raters rated the handwriting image display to be less than satisfactory. Early problems with connecting to the website and slow download speeds are nowadays disappearing as scanning technology and broadband speeds improve.

As late as 2003, Twing et al., reporting on essay marking in both mediums, found that some markers had never worked on a computer so, inevitably, some anxiety about marking on-screen occurred. Adams' (2005) study of OSM produced a mixed reaction, with raters not knowing how many scripts needed to be marked and how many remained. They also noted issues with returning to earlier scripts and marking continuously in front of a computer.

In summary, it can be seen that while some researchers have reported minor differences between the two methods of marking, in general, studies have reported data that suggest the two methods are largely comparable. Further, in most of the studies that have been conducted, raters have generally also been positive about OSM although some reservations have been expressed. Early problems with scanning technology and broadband speed are decreasing as both the technology and broadband speeds improve.

The data for the Hong Kong studies (see Coniam, 2009, forthcoming) involved 30 raters with good statistics (i.e., high inter-rater correlations and high correlations with the objectively-marked HKCE English Language Reading Paper 1A) from the 2007 HKCE Writing Paper rating on paper (i.e., re-rating) a set of scripts, most of which they had rated onscreen nine months previously. Subsequent to the rating, the 30 raters completed a questionnaire providing feedback on the exercise and participated in semi-structured interviews. It was reported that, technologically, raters had no problems with rating on screen. Statistical results suggested no bias favouring either form of rating from the correlations that emerged between the two forms of rating. Further, the number of discrepancy scripts (where a third rating is invoked because of a gap between the marks awarded by the two original raters of 5 or more marks out of a possible 24) was comparable with, if not lower than, that of the live 2007 HKCE Writing Paper (Coniam, 2009).

In the Coniam (2009) study, the two forms of rating were seen as being equivalent in terms of grades awarded to test takers. It was nonetheless observed in some cases that a number of quite large discrepancies existed for certain scripts in terms of the grades awarded to these scripts when rated on screen as compared to their grades when rated on paper. In comparing the two modes of rating, it emerged that whereas certain raters rated Script X on paper more than 5 marks higher (the re-rating trigger, see below) compared with the grade awarded by other raters to Script X on screen, the converse was true for other raters. In this opposite case, some raters awarded Script X a score 5 marks or more higher on screen than did other raters when rating on paper. It was therefore decided to undertake a study to investigate whether discrepancy scripts have defining features, whether the discrepancy scripts in the Coniam (2009) study were just 'rogue' scripts (i.e., scripts that receive anomalous grades from time to time for possibly unidentifiable reasons), or whether discrepancy scripts are a fact of life, an inevitable occurrence in a subjective rating situation where raters rate written scripts or oral performances.

The phenomenon of variation between raters' scores is not new; the variability between raters is the subject of considerable discussion in the literature. Among the reasons why test takers' scores may vary, McNamara (1996), for example, cites a range of causes: rater (mis)interpretation of the rating scales and descriptors; rater freshness (or tiredness); and interpersonal factors (albeit unintentional) where raters respond positively or negatively to certain gender, race, or personality types. Hamp-Lyons (1991) discusses the effect of the prompt; and Weigle (2002) discusses the different rating scales used (content, organisation etc) which direct raters in their assessment. Hamp-Lyons (1989) suggests that raters respond to cultural differences in writing, which may, in part, be attributable to their own cultural and experiential background. Vaughan (1991) notes that raters' reactions to different language features resulted in different grades being awarded to essays. Other factors which have long been recognised as affecting grades awarded to essays concern neatness and the quality of the handwriting (see e.g., Huck and Bounds, 1972; Chase, 2005)

One way of reducing variation involves rater training, the importance of which has long been accepted as an essential factor in a test's reliability (see e.g., Webb et al., 1990). While training is important in orienting raters towards the rating scale, it has been argued that it is not always possible to achieve very close levels of agreement between raters (Lunz and Stahl, 1990), and that even extensive training can have little effect on standards maintained by raters (Englehard, 1992; Weigle, 1998). Some researchers (e.g., Constable and Aldrich, 1984) have even argued whether perfect agreement is achievable or even desirable.

Webb et al. (1990), in discussing problems associated with rater stringency, leniency and inconsistency, state that while these issues may to an extent be mitigated by statistical adjustment, rater training is essential to solve other problems – specifically, rater inconsistency (see also Weigle, 1998). Lumley and McNamara (1995) state that if inconsistency is to be reduced, training and standardisation are not only essential, but that further moderation is required shortly before test administration because a time gap between the training and the assessment event reveals that inconsistencies re-emerge.

The variation between ratings in the study described above involved two raters giving significantly different scores in different rating mediums – OSM or PBM – not simply between raters in one medium, usually paper. Consequently, it was decided that a similar screen-versus-paper-based methodology would be utilised to investigate a set of scripts where significant differences in the ratings had occurred.

The criterion for invoking re-rating (the use of a third rater) has been long established for the HKCE Writing Paper as two raters differing from each other by more than one score point on a 6-point scale (see e.g., Attali & Burstein, 2005, p. 13). A comparable baseline exists for the 2007 HKCE Writing Paper, with the discrepancy rate between the two raters set at 5 points out of the 24 available. Using this criterion, the discrepancy rate for the 2007 HKCE Writing Paper was approximately 10% (HKEAA, personal communication regarding onscreen marking statistics, June 2007). The five-point difference is therefore taken as criterial in the current study.

The study

The data used in the study was drawn from Task 2 of the 2007 HKCE English language Writing Paper (candidature 99,771), for which test takers were required to produce a piece of expository writing of approximately 250 words from a choice of two prompts (Appendix 1). The first was a descriptive essay, where candidates had to explain why they would like to work in the fashion industry. The second was argumentative, with candidates having to argue whether it was more important to be clever than beautiful (HKEAA: 2007, p. 18). The HKCE Writing paper is rated via four subscales and descriptors, with each subscale having six levels – '6' indicating most, and '1' least able (HKEAA, 2007, pp. 104-106). All scripts are double rated, with a third rater invoked, as mentioned, where there is a discrepancy between the two raters of 5 or more out of the maximum of 24 points.

The study comprised 30 raters, each marking 100 scripts that they had rated previously – a total of 3,000 scripts. The script sample breaks down as follows. First, 24 scripts were used as 'control scripts'. They had been used in the original examination by the HKEAA for standardisation purposes. These were rated by all 30 raters. The 24 scripts also formed the data contact points for multi-faceted Rasch measurement (MFRM). Apart from the control scripts rated by all 30 raters, the remaining 76 scripts were rated by pairs of raters, giving a total of 2,280 paired ratings. The study therefore generated two sets of ratings:

1. 10,440 comparisons from the combinations of all 30 raters rating all control scripts. The formula below shows Rater X's rating for each control script compared against the other 29 raters, viz.:
([(30-1) x 30 raters ÷ 2 to remove duplicates from the two-dimensional matrix] x 24 scripts)

2. 2,280 ratings from the paired-rated scripts (30 raters x 76 scripts)

The most powerful factor in the current study is the set of control scripts. Since these were rated by multiple raters, the chances that certain scripts might receive differing grades was considerably greater than in scripts rated solely by one pair of raters. Table 1 summarises the discrepancy script situation across the two sets of scripts. It will be recalled that the discrepancy criterion is 5/24 or greater.

Double-rated scripts	2,280 paired ratings
Discrepancies	174/2,280 (7.6%)
Control scripts	10,440 paired ratings
Discrepancies	76/10,440 (0.72%)

Table 1: Discrepancies of 5 or greater

As can be seen from Table 1, the set of double-rated scripts recorded an overall discrepancy rate of 7.6% – a rather lower incidence of discrepancies than the 2007 HKCE Writing paper of 10%. The discrepancy rate for the control scripts was, unsurprisingly, much lower given that these had been specifically selected by chief examiners on the basis that these scripts represented subscale criterion levels. The overall discrepancy rate was 0.72%, with 76 rating pairs differing by 5 points or more.

A detailed examination of the ratings of the control scripts across the two mediums revealed that 15 of the 24 scripts had grades diverging by 5/24 points or more; i.e., scripts which had been rated by at least one rater at least five points *more severely on paper* than on screen, while at least one other rater had rated the same script at least five points *more severely on screen* than on paper.

As illustrated in the previous study (Coniam, 2009), Prompt 1 was more demanding than Prompt 2. Table 2 below now extends the analysis to provide an overall picture of the split by prompt for both double-rated scripts. As the situation in the HKCE English language Writing Paper involves examining the discrepancies between pairs of raters, the discussion will henceforth relate to the ratings produced between pairs of raters only – the upper part of Table 2.

	1. Fashion show	**2. Clever or beautiful**	**Total**
Double-rated scripts	534 (23.4%)	1,746 (76.8%)	2,280
Control scripts	9 (37.5%)	15 (62.5%)	24

Table 2. Prompt split

While the numbers were not identical, the prompt choice splits were not dissimilar. Approximately one quarter selected Prompt 1 in the double-rated scripts, with three quarters opting for Prompt 2. With the control scripts, the differential was rather narrower, being one-third to two thirds. This is not surprising as there were 2,280 different scripts double rated as against just 24 control scripts.

Having established the prompt split, Table 3 now presents the discrepancy figures for both prompts.

Prompt	Number of scripts	Discrepancy scripts
1. Fashion show	33/534 scripts (6.1%)	5 (33.3%)
2. Clever or beautiful	141/1,746 scripts (8.1%)	10 (66.7%)
Total	174/2,280 scripts (7.6%)	15 (100%)

Table 3. Discrepancies of +/- 5 (double-rated scripts)

As can be seen, a higher proportion of discrepancies emerged for Prompt 2, the more demanding of the two prompts. The split with regard to the set of 15 discrepancy scripts was 5 (33.3%) on Prompt 1, and 10 (66.7%) on Prompt 2, generally comparable with the overall figures presented for the two prompts in Table 2 above.

As a control against the 15 control scripts with contrasting discrepancies, a second set of scripts comprising 15 scripts was selected from the double-rated scripts, with the prompt split comparable to that of the control script set. There are, therefore, two contrasting groups, summarised in Table 4 below.

Group 1 consisted of control scripts selected from scripts with discrepancies +/- 5 or more between any two raters from the set of scripts rated by all 30 raters. Grades ranged from Level 2 (i.e., 8/24) to Level 5 (20/24); there were no Level 1s or 6s.

Group 2 scripts were selected on the basis of both raters having given exactly the same score to a script (see the detail in the research questions below) and rated by pairs of raters only. They also ranged from Level 2 (i.e., 5-8/24) to Level 5 (i.e., 17-20/24). There were no scripts with scores of 1 or 6. 15 such scripts were identified.

Group	Prompt	Discrepancy situation	Grade levels
Group 1 (101-115) Control scripts	#1 – 5/15 #2 – 10/15	All scripts discrepancies 5/24 or more between at least one pair of raters	Level 2 (i.e., 8/24) to Level 5 (20/24); no Level 1s or 6s.
Group 2 (201-215) Pair-rated scripts	#1 – 5/15 #2 – 10/15	Both scripts given exactly the same score by both raters	Level 2 to Level 5; no Level 1s or 6s.

Table 4. Two sets of scripts

Research questions

The major hypothesis in the current study is that discrepancy scripts will have identifiable features which differentiate them from scripts which have been awarded the same grade. The research questions investigating these features can be stated as:

1. Do discrepancy scripts produce greater misfit in the Rasch model than non-discrepancy scripts?

2. Do raters identify more discrepancy scripts as being 'problematic' to grade than non-discrepancy scripts?

3. Can raters pinpoint features that make the discrepancy scripts problematic for assessment?

Raters

12 raters participated in the current study. These were Cantonese speaking trainee teachers in their fourth and final year of a Bachelor of Education programme in English Language Teaching from a local university. They understand the English language capabilities of Hong Kong secondary students, all having attended Hong Kong primary and secondary schools and all having spent a number of periods of practice teaching in each of three years in different secondary schools. They were all proficient in English, operating at IELTS level 7. At the time of the study they were enrolled on a three-unit (39-hour) programme on Language Testing where current assessment issues were being explored. Oral testing, criterion-referenced assessment and the use of band scales had been recent topics.

As preparation for the assessment sessions, raters were first trained and standardised, following procedures adopted by the HKEAA for rater training. This involved raters first familiarising themselves with the scales and descriptors – although they were already familiar with these as final year trainee teachers – after which they rated a sample of 10 scripts at home. They then attended a half-day session of training and standardisation. First, feedback was taken on the sample scripts they had rated at home; this was followed by a further set of 10 more scripts for trial rating and comment. At the end of the training session, the 12 raters were assigned to two groups – Set 1 (Raters 1-6) and Set 2 (Raters 7-12).

A crossed rating design was implemented such that the six raters 1-6 rated the Group 1 scripts (101-115) on screen, while raters 7-12 rated the Group 2 scripts (201-215) on paper. The raters then changed around so that raters 1-6 rated Group 2 scripts on paper, with raters 7-12 rating the Group 1 scripts on screen.

As they rated, raters were asked to note whether any of the scripts were easy or problematic to rate. If they were, they wrote brief comments on the script in writing. After the exercise, raters took part in semi-structured interviews, each lasting for about 15 minutes, where – with reference to the notes that they made on the scripts, which they had to hand – they were asked to comment on the rating process and why it might have been problematic to rate certain scripts. Interviews were subsequently transcribed and analysed to identify salient points, i.e., common themes, identified and categorised across test takers. Semi-structured interviews were preferred to both structured interviews and stimulated recall because semi-structured interviews (see Rubin & Rubin, 1995) are conducted with a fairly open framework which allows for focused, conversational, two-way communication. In the semi-structured interview format, they are guided only in the sense that some form of interview guide is prepared beforehand, and provides a framework for the interview. The essence of the stimulated recall method (Bloom, 1953, p. 161) is that "the subject may be enable to relive an original situation with vividness and accuracy if he is presented with a large number of the cues or stimuli which occurred during the original situation". In this study, what was required was not a 'reliving' of the rating experience but a general impression of problems that the raters encountered in the scripts.

Data analysis

Test takers' final grades in Hong Kong English language public examinations are computed directly from raters' raw scores. While the latter may be adjusted for mean and standard deviation on the basis of correlations with other papers taken by the test takers, essentially the result is the raw score. The accuracy of information obtained from raw scores has long been questioned, with a number of studies commenting that the use of raw scores constitutes an imperfect measure of test taker ability (McNamara, 1996, p. 122; Weir and Shaw, 2008). A study by Coniam (2008), for example, examined the use of raw scores in the application of rating scales in the HKCE 2005 Writing Test and illustrated how the use of raw scores and measures derived through MFRM could produce markedly different results for test takers. Consequently, MFRM has been adopted as the statistic to be used in this analysis since it allows different facets to be modelled and their effects controlled.

In MFRM, the measurement scale derived by application of a unified metric such as the Rasch model means that various phenomena – rater severity-leniency levels, prompt difficulty, test taker ability etc – can be modelled (see McNamara, 1996, for an overview of the use of Rasch measurement in English language assessment).

In the current study, a five-faceted design was employed, modeling raters, test takers, input prompt materials, rating scales, and the rating medium. The computer program FACETS Version 3.61.0 (Linacre, 1994) was used to perform the analysis. The use of the Rasch model enables all these factors to be taken account of. First, in the standard Rasch model, the aim is to obtain a unified metric for measurement. This is not unlike measuring length using a ruler, with the units of measurement in Rasch analysis (referred to as logits) evenly spaced along the ruler. Logits are centered at zero, zero being the 50% probability represented by an "item" of average difficulty. Second, once a common metric is established for measuring different phenomena (test takers and test items being the most obvious), the phenomena can be examined and their effects controlled and compared. The use of a Rasch model of measurement provides, in principle, independence from situational features (the prompt, for example) in a particular test, with the results able to be interpreted with a more general meaning.

In MFRM, the measurement scale is based on the probability of occurrence of certain facets – in the current case, features associated with the rating of writing such as prompt difficulty, rater severity levels, and the rating medium. The phenomena – the different situational factors – can be explicitly taken into consideration and modelled in constructing the overall measurement picture.

FACETS reports model fit / misfit, as well as 'unexpected responses'. These two reports will form the basis for the analysis in the current study.

Results and discussion

This section first presents the result obtained from the MFRM analysis. It then moves to an analysis of raters comments on scripts as being easy or problematic to rate, followed by a synthesis of the semi-structured interviews.

Model fit / misfit

Overall data–model 'fit' – fit essentially being the difference between expected and observed scores – can be assessed by examining the responses that are unexpected given the assumptions of the model. According to Linacre (2004), satisfactory model fit is indicated when about 5% or less of (absolute) standardised residuals are equal or greater than 2, and about 1% or less of (absolute) when standardised residuals are equal or greater than 3.

In the current study, there were 1,532 valid responses used for estimating model parameters in the analysis for writing. Of these, 61 responses (3.98%) were associated with (absolute) standardised residuals equal or greater than 2, with 1 response (0.07%) being associated with (absolute) standardised residuals equal or greater than 3. These findings, along with the fit statistics for the different facets (presented below) suggest satisfactory model fit.

To give the overall picture of facet placement, Figure 1 below presents the variable map produced by the computer program FACETS representing the calibrations of the five facets – raters, test takers, prompts, rating method, and the four rating subscales used to score test takers – with the different facets' location on the map, or vertical ruler.

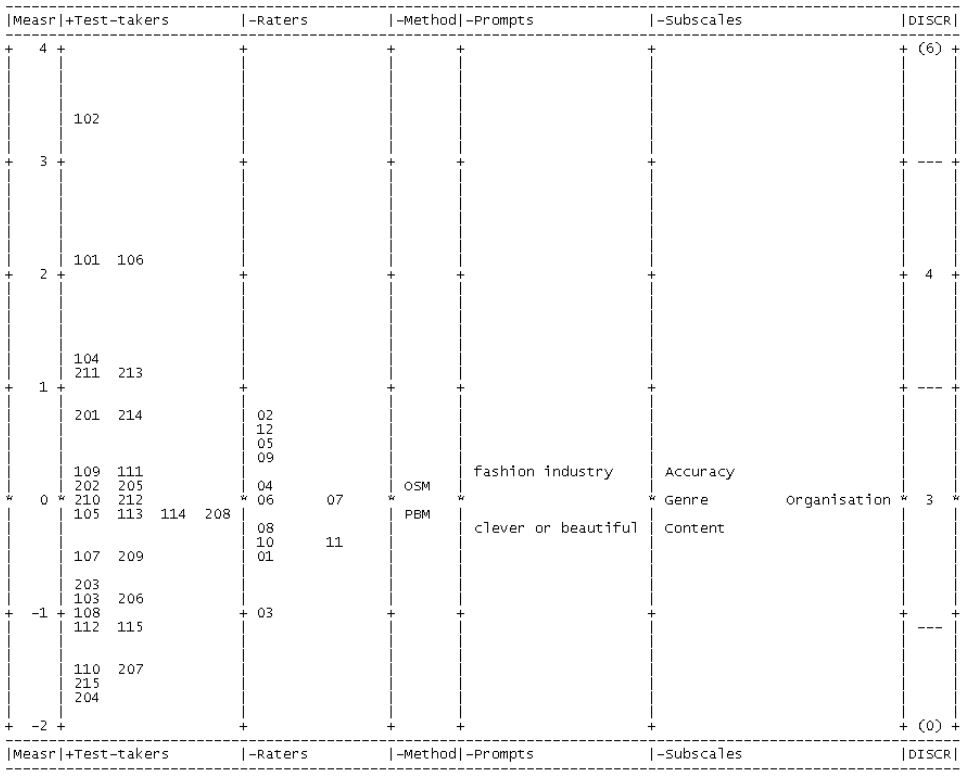

Figure 1: The five facets on FACETS' vertical ruler

As can be seen from Figure 1, the 30 test takers show some spread of ability, ranging from -2 to +1 logits. A spread was not necessarily aimed at, with test takers' selected solely on the basis that their essays were classified as discrepancy scripts. While the raters show a degree of spread, the other three facets are more closely clustered around the zero logit mark.

Since raters' internal consistency is one of the cornerstones in a test of writing, Table 5 below presents the results for raters.

Measure	Model error	Infit mean square	Raters
-0.32	0.13	1.43	10
+0.54	0.13	1.35	05
+0.02	0.13	1.24	06
+0.67	0.13	1.09	12
-0.05	0.13	1.08	07
-0.27	0.13	1.05	08
+0.77	0.13	0.98	02
-0.37	0.13	0.91	11
+0.14	0.13	0.83	04
-0.98	0.13	0.74	03
-0.53	0.13	0.68	01
+0.38	0.13	0.57	09
0.00	0.13	1.00	Mean
0.53	0.00	0.26	S.D.

Separation 3.92, Reliability 0.94, chi-square 180.0, d.f. 11, significance .00

Table 5. Raters' measurement report

In Table 5, Column 3 presents the infit mean square statistic, which describes model fit, for which there are different interpretations. 'Perfect fit' according to Bond and Fox (2007, pp. 285-286) is defined as 1.0, with an acceptable upper limit of fit stated as 1.3. Weigle (1998) proposes acceptable practical limits of fit as 0.5 for the lower limit and 1.5 for the upper limit. Given this, we can see that with the exception of Raters 10 and 05, 10 of the 12 raters show good fit. The rater logit range is from +0.77 to -0.98, a comparatively narrow range of 1.75 logits. While figures for rater range vary, a range of under 3 logits shows a comparatively narrow spread compared to other studies involving the rating of writing. 3.42 logits was recorded in the Coniam (2008) study, with a 4.26 logit spread reported in Eckes (2005). The reliability of 0.94 indicates that raters are being reliably separated into different levels of severity and the chi-square figure indicates that significant differences in leniency among raters exist, with the null hypothesis thus rejected.

To provide a more complete picture, Tables 6 and 7 present the results for the prompts and the rating methods.

Measure	Model error	Infit mean square	Prompt
+0.30	0.07	1.11	Fashion industry
-0.30	0.04	0.95	Clever or beautiful
0.00	0.06	1.03	Mean
0.30	0.01	0.08	S.D.

Separation 5.22, Reliability .96, chi-square 56.6, d.f. 1, significance .00

Table 6. Prompt measurement report

Both prompts fit the model well. Prompt 1 at +0.31 logits (a descriptive essay requiring test takers to explain why they would like to work in the fashion industry) emerged as slightly more demanding than Prompt 2 at -0.31 logits (an argumentative essay requiring test takers to put the case for whether it was more important to be clever than beautiful). As noted in Coniam (2009), the fact that Prompt 1 emerges as more demanding is perhaps not surprising, as the vocabulary needed for arguing about being clever or beautiful will be more accessible to test takers than will the schema regarding fashion or working in the industry, given that test takers are generally 17 years old and not in full time employment. Results were almost identical with those of the Coniam (forthcoming) study, where Prompt 1 was +0.31 logits and Prompt 2 was -0.31 logits. The chi-square value indicates significant differences between the difficulty levels of the two prompts.

Measure	Model error	Infit mean square	Rating method
+0.13	0.05	1.01	OSM
-0.13	0.05	0.98	PBM
0.00	0.05	1.00	Mean
0.13	0.00	0.01	S.D.

Separation 2.24, Reliability .83, chi-square 12.0, d.f. 1, significance .00

Table 7. Rating method measurement report

With regard to the method of rating, there was a slightly larger differential in the logit values compared with the Coniam (forthcoming) study. While the rating methods both exhibited good fit, the chi square figure is nonetheless indicative that there are significant differences between the two methods of marking in the current study.

Table 8 below presents an analysis of test takers.

Measure	Model error	Infit mean square	Test takers
+1.09	0.21	1.50	211
-1.16	0.22	1.36	112
-0.07	0.21	1.33	114
+2.14	0.21	1.22	101
+0.28	0.21	1.13	111
-1.51	0.21	1.13	207
+0.15	0.21	1.09	202
-0.08	0.15	1.09	105
+2.11	0.21	1.07	106
-0.88	0.21	1.07	206
-1.45	0.21	1.07	110
+1.22	0.15	1.05	104
+1.16	0.21	1.05	213
-1.02	0.21	1.03	108
-0.49	0.21	1.02	209
+0.23	0.21	0.97	109
-0.54	0.21	0.94	107
-0.75	0.21	0.94	203
-0.15	0.21	0.93	208
+0.06	0.21	0.89	205
+0.74	0.20	0.87	201
+0.74	0.20	0.86	214
+0.02	0.21	0.86	212
-1.10	0.21	0.83	115
-0.88	0.21	0.78	103
-0.15	0.21	0.77	113
+0.02	0.21	0.74	210
-1.77	0.21	0.72	204
-1.62	0.21	0.70	215
+3.41	0.25	0.67	102
-0.01	0.21	0.99	Mean
+1.21	0.02	0.20	S.D.

Separation 5.78, Reliability 0.97, chi-square 926.8, d.f. 29, significance .00

Table 8: Test takers' measurement report

As mentioned earlier, with the exception of a couple of outliers, most test takers were clustered in a 3-logit range. It had been expected that there might be more misfit, despite the fact that half of the data set contained subjects whose scores were substantially different. This was not the case, however, and with the exception of one or two test takers, most fit the model acceptably.

It had been anticipated that many of Group 1 test takers – the discrepancy test takers (i.e., #101-#115) – would show misfit. This was not the case. Indeed, of the three test takers who exhibited misfit, the most misfitting (#211) was from Group 2, the same-grade script group.

FACETS 'unexpected response' data

FACETS reports 'unexpected responses'. While this, to an extent, supports the fit data – or rather the data which does not fit – in the current study this is a source of possible clues as to what fits and what does not.

Table 9 below therefore gives an indication of how the 65 unexpected responses were reported across the different facets.

Rating method	Prompt	Rater	Script type
OSM = 33 PBM = 32	#2 (Clever) = 39 #1 (Fashion) = 26	Rater 05 = 12 Rater 10 = 9	Group 1 (discrepancy scripts) = 36 Group 2 (same-grade scripts) = 29

Table 9. FACETS unexpected responses (N=65)

As can be see from Table 9, it would appear that there are no clear indicators of irregularity. The rating method presents an approximately even split. Likewise, the split for the prompt is also approximately even, given that the distribution of unexpected responses mirrors the popularity of the prompt. The raters with most unexpected responses were those with the worst infit statistics – Raters 05 and 10 (see Table 5). This is not wholly surprising as it could be surmised that 'poor' ratings might give rise to more 'unexpected responses'. Group 1 recorded 36/65 (55.3%) unexpected responses while Group 2 recorded a slightly lower amount at 29/65 (44.7%). In the final analysis, both types of rating gave rise to almost identical amounts of unexpected responses, although there were in fact slightly more discrepancy scripts recording unexpected responses than in the same-grade scripts. To summarise, neither the rating method nor the prompt made a difference to the results. The rater was a factor but that is probably because the two raters had poor rater statistics anyway. There was a difference between Group 1 and 2 results but this is only one facet. Further research related to these issues needs to be undertaken.

Scripts being problematic or easy to rate

As they rated their two sets of 15 scripts, raters were asked, if they considered it appropriate, to comment as to whether particular scripts were easy or problematic to rate, and to comment why. Tables 10a and 10b summarise their opinions.

Test taker	Easy to rate	Problematic to rate	No comment
101	2	2	8
102	4	0	8
103	2	2	8
104	7	1	4
105	5	3	4
106	3	1	8
107	2	4	6
108	2	3	7
109	1	2	9
110	2	1	9
111	2	1	9
112*	3	0	9
113	3	0	9
114*	3	0	9
115	2	2	8
Totals	43 (23.9%)	22 (12.2%)	115 (63.9%)

Table 10a. Group 1 – Discrepancy scripts (* = poor model fit)

Test taker	**Easy to rate**	**Problematic to rate**	**No comment**
201	4	3	5
202	4	2	6
203	3	2	7
204	**1**	**5**	**6**
205	4	1	7
206	3	2	7
207	5	1	6
208	5	1	6
209	5	1	6
210	3	3	6
211*	3	2	7
212	**2**	**5**	**5**
213	4	2	6
214	4	2	6
215	**2**	**7**	**3**
Totals	52 (28.9%)	39 (21.7%)	89 (49.4%)

Table 10b. Group 2 – Same-grade scripts
(Bold font = numerous problematic comments)

F`rom the maximum of 180 possible comments per group (15 scripts x 12 raters), both groups recorded more 'easy to rate' than 'problematic to rate' comments. Scripts were, on average, commented on by a third to a half of the raters. On scripts being easy to rate, 43 comments (23.9%) were received for Group 1, with 52 (28.9%) for Group 2. With regard to scripts being problematic to rate, 22 comments (12.2%) were received for Group 1, with 39 (21.7%) for Group 2. On balance, it can be seen that, contrary to expectation, raters generally reported scripts as being easier to rate than problematic, although both types were apparently evenly spread across the two groups of scripts. While in Group 2 a higher number of scripts was considered easy to rate, almost the same number of scripts was considered problematic – almost double that of Group 1. It will be recalled that test takers who showed most misfit in the Rasch model were #112 and #114 in Group 1 and #211 in Group 2 (asterisked in Tables 10a and 10b above). Group 1 test takers #112 and #114 received no problematic comments, with #211 receiving more 'easy to rate' comments than 'problematic to rate' ones. All the test takers who received most 'problematic' comments (in bold font in Table 10b above) were from Group 2. These were test takers #215, with 7 comments, and #204 and #212 with 5 comments. None of the three test takers, however, showed poor Rasch model fit. In line with the problematic rating, raters made a number of comments on test takers #215, #204 and #212; these are summarised in Table 11.

Test taker	Comments
#215	• Terrible handwriting • Words too packed together
#204	• Arrows everywhere [indicating text to be moved] • Too much error correction fluid used • It's more like a draft • Students' words have to be traced around the page from time to time
#212	• Too many corrections • Many inserted words are too small to read

Table 11. Comments on 'problematic' scripts

As Table 11 illustrates, raters essentially commented on issues related to 'legibility', such as poor handwriting, overuse of correction fluid and generally messy appearance in terms of corrections and insertions. As it happened, however, all these comments related to three scripts from Group 2. So while the issue is one which irks raters, in the current study it has not emerged as being a feature of the (Group 1) discrepancy scripts.

Qualitative feedback

The last piece of data relates to the semi-structured interviews held with the 12 raters. From these interviews, five major categories emerged from an analysis of the common themes in the raters' comments. These are presented in Table 12.

Comment	**Raters commenting**
1. Use of correction fluid by test takers	7
2. Quality of the scanned scripts with regard to their general onscreen readability	5
3. Certain test takers' poor handwriting	8
4. Scrolling being problematic, hindering the ability (or desire) to reread a script	8
5. The *Genre* subscale being difficult to interpret appropriately	5

Table 12: Categories emerging from the interviews

The first four issues concern what might be referred to as 'construct irrelevant variance'. As an elaboration of the above Table, some of the raters' comments will now be reported. The fifth issue above, which will not be discussed further since it is outside the scope of the current study, relates to rater understanding of the rating scales and hence the issue of standardisation.

1. The use of correction fluid

01	"The use of correction fluid terribly affected the quality of the scanned version … the correction fluid mark affected me reading the words."
06	"When marking on screen, I came across some difficulties if the students used correction fluid in the scripts".

2. The quality of the scanned scripts with regard to onscreen readability

12	"After scanning, the images became even harder to be read."
02	"It was clearer to read the words on paper."

3. Test takers' poor handwriting

04	"Some students handwriting was illegible on screen ... but I found it could be easily read or guessed on paper."
08	"Some of the handwriting looked very messy on screen and looking at the screen for a long time was very tiring."

4. Scrolling being problematic

07	"I prefer to mark on paper really as looking at the "physical" script is better, my eyes got much less tired. Also when I did the marking onscreen, I just didn't bother at times as I didn't want to scroll up and down and make my eyes even more tired."
11	"I could flip over the pages easily when I wanted to refer to a particular page for comparison. It was a bit troublesome and inconvenient to move the mouse up and down when I wanted to look at the previous page."

Conclusion

As Hong Kong moves towards universal OSM for its national examinations, the current study has attempted to investigate the factors that cause concern for raters involved in the grading of 'discrepancy scripts', i.e. scripts that show such disparities in the grades awarded by two separate raters that the need for a third rater is triggered. From the previous study Coniam (2009), a set of 15 scripts were identified across the two rating mediums which had grades diverging by 5/24 points or more; i.e., scripts which had been rated by at least one rater 5 points more severely on paper than on screen, while at least one other rater had rated the same script 5 points more severely on screen than they had on paper. As a control, a second set of scripts comprising 15 scripts was selected from the double-rated scripts. These scripts had received exactly the same grade from two different raters. The two groups of scripts mirrored as far as possible the prompt split in the larger study and encompassed a range of ability levels – from Level 2 to Level 5.

Most data fit the Rasch model acceptably. An examination of test takers showed that, with the exception of three test takers – two from Group 1 and one from Group 2 – all data showed good model fit; the expected plethora of misfit from the Group 1 test takers did not emerge.

With regard to raters reporting scripts as being easy or problematic to rate, more comments were received on both counts for Group 2 than for Group 1. Further, contrary to expectation, whereas Group 2 recorded approximately equal numbers of easy and problematic scripts, Group 1 recorded twice as many scripts as being easy to rate rather than problematic.

Finally, in the semi-structured interviews, four major issues were raised by a number of raters with regard to problems in rating scripts. These were the use of correction fluid; general onscreen readability of the scanned scripts; poor handwriting; and scrolling being problematic. These issues, however, have been found in other studies and are not specific to the Group 1 scripts alone (see Whetton & Newton (2002), for example). It is likely that these issues will diminish in time as raters grow increasingly familiar with OSM and as script scanning technology develops.

In conclusion, the major hypothesis that discrepancy scripts have identifiable features has to be rejected as have the three subsidiary questions arising from it. It may well be a fact of life that there will be discrepancies – occasionally large ones – between raters. This may be a disappointing conclusion but, as discussed in the opening section of this study, researchers such as Constable and Aldrich (1984) have suggested that some rater variation may be inevitable. The current research underscores this point. The major approach to dealing with and mitigating this variation is continual training and standardisation. In addition, in measurement terms, understanding and monitoring how test taker grades are arrived at by modeling the different facets via the use of a statistic such as multi-faceted Rasch measurement allows for the effects of the different facets to be controlled. However, even if these twin measures are implemented, it is unlikely that discrepancy scripts can be obliterated completely so it is likely that examination agencies will have to accept the inevitability of their occurrence.

Acknowledgement

I would like to thank the Hong Kong Examinations and Assessment Authority – and in particular Christina Lee, the General Manager for Assessment Development – for support on the project: for access to raters' scores and to test takers' scripts and data.

References

Adams, C. (2005). How does assessment differ when e-marking replaces paper-based marking? Paper presented at the 31st Annual International Association for Educational Assessment Conference "Assessment and the Future of Schooling and Learning". September 4-9 2005, Abuja, Nigeria.

Attali, Y., & Burstein, J. (2005). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, 4(3). Retrieved May 20, 2008, from http://escholarship.bc.edu/jtla/vol4/3/.

Bloom, B.S. (1953). Thought-processes in lectures and discussions. *Journal of General Education*, 7, 160-169.

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*, 2nd edition. Mahwah, N.J.: Lawrence Erlbaum.

Chase, C. I. (2005). Essay test scoring: Interaction of relevant variables. *Journal of Educational Measurement*, 23 (1), 33-41.

Coniam, D. (2008). An investigation into the effect of raw scores in determining grades in a public examination of writing. *Japan Association for Language Teaching Journal*, 30 (1), 69-84.

Coniam, D. (2009). A comparison of onscreen and paper-based marking in the Hong Kong public examination system. *Educational Research and Evaluation,* 1 (3), 243- 263

Coniam, D. (forthcoming). Validating onscreen marking in Hong Kong. *Asia Pacific Education Review*.

Constable, E. & Andrich, D. (1984). Inter-judge reliability: Is complete agreement among judges the ideal? Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 24-26, 1984).

Eckes, T. 2005. Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, 2 (3), 197 - 221.

Englehard, G. Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, 5, 171-191.

Hamp-Lyons, L. (1989). Second language writing: assessment issues. In Kroll, B. (ed.) *Second language writing* (pp. 69-87). Cambridge: Cambridge University Press.

Hamp-Lyons, L. (1991). Pre-text: task-related influences on the writer. In Hamp-Lyons, L. (ed.) Assessing second language writing in academic contexts (pp. 87-107). Norwood, NJ: Ablex.

Huck, S. W. and Bounds, W. G. (1972). Essay grades: An interaction between graders' handwriting clarity and the neatness of examination papers. *American Educational Research Journal*, 9, 279-283.

Linacre, J. M. (1994). FACETS Version 3.61.0: Rasch Measurement Computer Program. Chicago: MESA Press.

Linacre, J. M. (2004). A user's guide to FACETS: Rasch-model computer programs [Software manual]. Chicago: Winsteps.com

Lumley, T. & McNamara, T. (1995). Rater Characteristics and Rater Bias: Implications for Training. Language Testing, 12(1), 54-71.

Lunz, M. & Stahl, J. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Profession*, 13, 425-444.

McNamara, T. (1996). *Measuring second language performance*. New York: Longman.

Newton, P., Whetton, C. Adams, E. Bradshaw, J. & Wong, C. (2001) An Evaluation of the 2001 New Technologies Pilot, NFER.

Powers, D. & Farnum, M (1997) Effects of mode of presentation on essay scores. ETS Report RM-97-08.

Powers, D., Farnum, M., Grant, M & Kubota, M (1998). Qualifying essay readers for an on-line scoring network. ETS Report RM-98-20.

Powers, D., Kubota, M., Bentley, J., Farnum, M., Swartz, R. & Willard, A. (1997). A pilot test of on-line essay scoring. ETS Report RM-97-07.

Rubin H. J. & Rubin I. S. (1995) *Qualitative Interviewing: The Art of Hearing Data*. London: Sage.

Sturman, L. & Kispal, A. (2003). To e or not to e? A comparison of electronic marking and paper-based marking. Paper presented at the IAEA Conference, Manchester, Retrieved 20 June 2007 from http://www.aqa.org.uk/support/iaea/papers.html.

Twing, J., Nichols, P. & Harrison, I. (2003). The comparability of paper-based and image-based marking of a high-stakes, large-scale writing assessment. Paper presented at the IAEA Conference, Manchester, Available at: http://www.aqa.org.uk/support/iaea/papers.html

Vaughan,C. (1991). Holistic assessment: What goes on in the writer's mind? In Hamp-Lyons, L. (ed.) Assessing second language writing in academic contexts (pp. 111-125). Norwood, NJ: Ablex.

Webb, L., M. Raymond & Houston, W. (1990). Rater Stringency and Consistency in Performance Assessment. Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing* 15 (2), 263-287.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Weir, C. and Shaw, S. (2008). A socio-cognitive approach to writing test validation. In Taylor, L. and Weir, C. (eds.) Multilingualism and assessment, pp. 147-156. Cambridge: Cambridge University Press.

Whetton, C. & Newton, P. (2002,). An evaluation of on-line marking. Paper presented at the 28th International Association for Educational Assessment Conference, Hong Kong SAR, China, September 2002.

Zhang, Y., Powers, D., Wright, W. & Morgan, R. (2003). Applying the online scoring network (OSN) to advanced program placement program (AP) Tests. ETS Research Report RR-03-12.

Appendix 1: 2007 HKCE English Language Writing Paper, Task 2

Write about 250 words on ONE of the following topics.

1. You would like to enter the essay competition advertised in the poster below. Read the poster and write your essay.

Win 6 weeks' work experience in the fashion industry.

Would you like to work
with a famous fashion designer;
on a popular fashion magazine;
OR
in a shop selling very expensive clothes?

Choose ONE of the above and write an essay explaining the reasons for your choice.
Email your essay to essay@hkfashion.com
Entry deadline: Friday 4th May, 2007

2. 'It is more important to be clever than beautiful or handsome.' Do you agree.

Write a letter to the editor of the Young Post giving your opinions. Start your letter 'Dear Editor', and sign it 'Chris Wong'. Do not write an address.