Education Journal《教育學報》, *Vol. 37, Nos. 1–2, Summer–Winter 2009, 71–87*

Examining Negative Attitudes Toward Onscreen Marking in Hong Kong

David CONIAM
Department of Curriculum and Instruction
The Chinese University of Hong Kong

This article details an investigation into onscreen marking (OSM) in Hong Kong — where paper-based marking (PBM) is being phased out, to be completely superseded by OSM. It is a specific follow-up to a larger study (Coniam, 2009a) involving 30 raters who had previously rated English language essay scripts on screen in the 2007 Hong Kong Certificate of Education Examination (HKCEE). In that study, 16 raters were generally negative about marking onscreen compared with marking on paper as against 8 raters who were generally positive about OSM. The current study is a direct response to concerns that the attitudes of the two groups of raters (i.e., negative versus positive attitude) might be reflected in the scores awarded to test takers through the two marking mediums. An examination of the groups' data involving classical measurement statistics results such as correlations between rater attitude and the different component of the HKCEE Writing paper, along with multi-faceted Rasch measurement to examine rater fit and erratic behavior in marking, reveals that a negative attitude toward OSM does not appear to impact upon the reliability of the rating.

Paper-based marking (PBM) is to be phased out in Hong Kong and replaced totally in 2012 by onscreen marking (OSM) in public examinations. In order to investigate and validate the adoption of OSM in Hong Kong, a series of studies has been conducted to compare the two methods of marking, the first of these being Coniam (2009a). These have been carried out in the context of a pilot examination — one of the Year 11 (Secondary 5) English language public examinations, for which

OSM was adopted as the sole method of marking in 2007. To orient the reader, a brief overview of the previous study, along with the results that emerged from it will be presented below. For the full background and major details, however, the reader is referred to Coniam (2009a).

A number of studies in the late 1990s were conducted in the United States by the Educational Testing Service (ETS) on different aspects of OSM (Powers & Farnum, 1997; Powers, Farnum, Grant, & Kubota, 1997; Powers, Kubota, et al., 1998). These studies suggested that scores were not affected by the medium in which essays were presented to raters, on screen or on paper. These findings were corroborated by further studies in the United Kingdom (U.K.) where Whetton and Newton (2002) evaluated online marking and, employing both expert and non-expert raters, reported little difference in the overall ratings of either group, although there were some differences in results between writing and spelling/handwriting. In addition, Sturman and Kispal (2003), investigating reading, writing and spelling tests, reported no consistent trends in the differences in test scores between the two methods of rating. Twing, Nichols, and Harrison's (2003) study, however, which compared the marking of essays on screen and on paper, concluded that the paper-based system was slightly more reliable than the onscreen one.

In a further ETS study, Zhang, Powers, Wright, and Morgan (2003) compared OSM (over the Internet) with PBM and found that while there were statistically significant differences between the mean scores from each method of scoring, the differences appeared equally likely to favor OSM as PBM. Regarding inter-rater reliability, no statistically significant difference emerged between the two methods of scoring, leading Zhang et al. to conclude that results obtained from OSM methods could be seen to be comparable to those obtained through PBM methods.

The adoption of OSM has not been accepted totally, however, as a degree of uncertainty surrounds certain aspects of OSM. Fowles and Adams (2005), with reference to the Assessment and Qualifications Alliance's e-marking experience in the U.K., concludes that OSM is different from PBM. As a result, she voices possible concerns about the validity and reliability of assessments made under OSM. In general, while feeling that current evidence supports the case for OSM, she nonetheless calls for further research to be conducted, with a "cautious approach" needed to ensure that important stakeholders, in particular

governments and teachers, are comfortable with the changes demanded by OSM. The series of studies being conducted in the Hong Kong context reflect these preoccupations.

In the U.K., after piloting OSM in 2006, Cambridge Assessment decided to invest substantial sums into OSM for the five-year period up to 2012 (Raikes, Greatorex, & Shaw, 2004). Further use of OSM has occurred in the Chinese mainland where more than 20 provinces have been practicing a limited form of OSM for a number of years. The Hong Kong Examinations and Assessment Authority (HKEAA) has been investigating computerization of various procedures and processes related to examinations for some time. After investigations and feasibility studies in 2005 into scanning facilities and the establishment of dedicated OSM centers, the Hong Kong Special Administrative Region's Legislative Council, in December 2005, allocated approximately US$25 million toward the modernization of information technology for the HKEAA, including the implementation of OSM (Legislative Council Finance Committee, 2005). Three special OSM centers were consequently established in strategic locations around Hong Kong, with a total of about 1,000 marking-dedicated workstations ready for the wholesale implementation of OSM in 2012. For a more in-depth picture of the actual implementation of OSM in Hong Kong in 2007, the reader is referred to Coniam (2009a).

While small-scale marking studies have been conducted in different countries, Hong Kong is the first jurisdiction to implement OSM across its entire public examination system. From 2012 onward, all scripts for all subjects will be marked using OSM. This quantum shift in marking procedures has, therefore, provided the impetus for initiating the current series of validation studies — with the implication that the full implementation of OSM in Hong Kong will have relevance and significance for other countries and jurisdictions.

The Current Study

The current study builds on, and elaborates, the findings of a previous study (Coniam, 2009a). The details of this study was presented to the HKEAA Research and Development Committee[1] in early March 2009. The Committee accepted the report, but raised the issue of Hong Kong teacher markers' attitudes toward the general adoption of OSM (details below). The Committee had concerns about the attitude of disaffected

markers and wondered whether those attitudes might be reflected in the scores awarded to test takers. These concerns provided the impetus for the current study as will be outlined in more depth below — i.e., the hypothesis that markers poorly disposed toward onscreen rating are likely to mark more severely or erratically.

Since the current study builds on a previous study, the previous study and its major findings will be briefly described in order to orient the reader. For background details of the Hong Kong education and examination system, the reader is referred to Coniam (2009a).

Background to the Previous Study

The data used in the previous and current study was drawn from the English Language Writing Paper Task 2 of the 2007 Hong Kong Certificate of Education Examination (HKCEE) (candidature just under 100,000), where test takers were required to produce a piece of expository writing of approximately 250 words (see Appendix). Test takers had a choice of two prompts — one descriptive, one argumentative (HKEAA, 2007, p. 18). The HKCEE Writing paper is rated via four subscales and descriptors; each subscale has six levels — "6" indicating most, and "1" least able (HKEAA, 2007, pp. 104–106). All scripts are double rated. A third rater is invoked where a discrepancy occurs between the two raters of 5 or more points out of the maximum of 24 points.

Of the 196 raters who marked the 2007 HKCEE Writing paper, 46 were identified as potential raters for the study on the basis of two criteria: (a) raters had good marking statistics in their rating of the 2007 HKCEE Writing paper (generally meaning an inter-rater correlation of above 0.8); (b) as far as possible, the sample would be a representative cross-section of raters with regard to gender and qualifications, teaching and rating experience. In the 2007 examination, 117 (59.7%) of the 196 raters were experienced raters. The remaining 79 (40.3%) were first-time raters. In the Coniam (2009a) study, efforts were made to recruit a number of first-time raters who had not marked by the traditional paper-based method and whose first-time marking experience would have involved the new status quo (i.e., OSM). This would provide an interesting point of comparison between experienced raters who had always marked on paper (and for whom OSM was a new experience),

compared with first-time raters who would only have rated on screen, and whose "new" experience would, conversely, be PBM.

Of the 46 potential participants, 30 raters were eventually recruited to take part in the study — 5 (16.7%) new and 25 (83.3%) experienced raters, with each rater marking 100 scripts. While it would have been preferable for the distribution of new/experienced raters in the sample to match the distribution in the live HKCEE Writing paper, it was not possible to achieve this for two reasons: (a) more experienced raters had better rating statistics and were thus more "eligible" to participate; (b) fewer new raters expressed an interest in participating.

When they were recruited for the study, the raters were informed that they would be marking scripts from the 2007 HKCEE, and that their batch of 100 scripts would contain some of the scripts they had marked in the previous year. They were not informed that they would in fact be re-marking 100 scripts from the batches they had previously marked in the 2007 examination. This procedure has been used successfully before (Coniam, 1991); the time lag of nine months is sufficiently long for raters not to recollect having marked the scripts before, rendering them as unfamiliar as unseen scripts (see Cheng, 1993). With 30 raters participating in the study, and each rater marking 100 scripts, the total sample therefore comprised 3,000 scripts, of which there were 2,145 different test takers. Scripts were also carefully selected from each marker's batch to cover the full range of levels (i.e., 1 to 6) of the subscales.

The study had two major research questions. The first research question hypothesized that there would not be statistical comparability between the two marking mediums. The hypothesis was not proven as comparable inter-rater reliabilities emerged between raters marking in the two different mediums, and test takers were found to have received comparable scores in either medium. The second question investigated raters' attitudes toward OSM, given the long history of PBM in Hong Kong. While raters were found to possess adequate technological competence to operate within the new OSM medium, attitudes to OSM were generally more negative than positive. While new first-time raters were, on the whole, positive about OSM, many of the experienced raters, who had long marked on paper, expressed reservations about the OSM process and its implications.

In the main study (Coniam, 2009a), raters completed a questionnaire detailing their attitudes toward the OSM and PBM processes. In terms

of attitudes between the two sets of raters — old and new, there were clear differences. In terms of general preference for OSM versus PBM, new raters rated OSM more positively than experienced raters. Also, in terms of a preference for marking at home or at a center, new raters rated center marking much more positively than did experienced raters; new raters also felt that having to travel to a special marking center was less of an inconvenience than did experienced raters. The stated hypothesis that "raters will not be negative about the OSM medium, showing no preference for either marking medium" could therefore be neither proved nor disproved, although the tendency was for a more negative then positive orientation.

A number of follow-up studies are being conducted to further validate the OSM process — extending the research questions from the previous study with regard to the statistical veracity of the rating and raters' attitudes. One of these, a qualitative study currently in progress, involves semi-structured interviews with a sample of the raters in order to discover whether "thick" data can provide further insights into rater attitudes and behavior (Geertz, 1973).

The current study extends the second research question from the original study and investigates whether raters' attitudes affect the OSM mark awarded. Specifically, it pursues two hypotheses:

1. Raters who hold a negative attitude toward OSM will rate test takers more harshly than will raters who have a positive attitude.
2. Raters who hold a negative attitude toward OSM will be more erratic than more positively oriented raters.

Data

As mentioned above, after rating, raters completed a questionnaire detailing their attitudes. Items were set on a 6-point Likert scale where "6" essentially denoted a positive response and "1" a negative response. There were 22 items relating to participants' view of their technological prowess and attitude toward the OSM process. The alpha for these 22 items was 0.85, indicating good reliability. In addition to the Likert-scale items, participants were asked to provide open-ended comments. Of the 26 raters who provided substantive comments, 16 were generally negative, 8 were generally positive, while 2 were neutral.

The correlation between the total score of the 22 items and the tenor of the attitude expressed in the open-ended question was 0.48 ($p < .01$), with the moderate correlation (Burns, 2000, p. 235) at the 1% level indicating that the two variables are tapping the same construct of negativity. Since the open-ended comments revealed a more overt expression of attitude, it is this variable that will be used to group raters. On this basis, the analyses conducted below will involve comparing the group of 16 raters with a negative attitude against the group of 8 with a positive attitude. The analyses conducted will build on those in the previous studies, namely an examination of scores obtained from the two groups of raters.

For Hypothesis 1, classical measurement statistics such as correlations of the two groups' scores on the Writing paper against the other components of the HKCEE will be adopted — as in the Coniam (2009a) study, and in line with standard HKEAA practice for determining reliability.

For Hypothesis 2, fit results derived from multi-faceted Rasch measurement will be presented, as in Coniam (2009b).

Results and Discussion

In the following section, the first issue discussed concerns scores awarded to test takers by the negative and positive attitude raters. This is followed by an examination of rater erraticness.

Scores Awarded to Test Takers

In the Coniam (2009a) study, one of the first issues analyzed — although a background variable — was the prompt. There were two prompts in the 2007 HKCEE Writing paper: Prompt 1 was descriptive, requiring test takers to write about working in the fashion industry for a week; Prompt 2 was argumentative, requiring test takers to discuss the pros and cons of being clever or beautiful. In the Coniam (2009a) study, Prompt 2 emerged as significantly more demanding than Prompt 1. Table 1 reproduces the prompt difficulty figures from the previous study.

While Prompt 2 was more demanding than Prompt 1, the results from *t* tests run with the marking method (OSM vs. PBM) as a grouping variable showed no significant difference between either method with

very comparable mean scores obtained from both methods. As Table 1 illustrates, the mean scores obtained by test takers (out of the possible maximum of 24) on Prompt 1 was 11.48 for OSM and 11.41 for PBM. For Prompt 2, the figures were 13.35 and 13.38 for the two marking mediums respectively. Thus, although the prompts affected mean scores, the medium used did not.

Table 1: Prompt Difficulty by Marking Method

Prompt	Marking method	*N*	*M* (max. 24)	*SD*	*t* test results
1. Working in fashion industry	OSM	810	11.48	5.32	*ns*
	PBM	810	11.41	5.38	
2. Clever or beautiful	OSM	2,190	13.35	5.13	*ns*
	PBM	2,190	13.38	5.24	

Rater Attitude as a Grouping Variable

Table 2 now lays out the data obtained by grouping the 16 negatively oriented raters (scripts marked = 1,600) and the positively oriented raters together (scripts marked = 800). *T* tests were then conducted for each prompt, with attitude as the group variable.

As can be seen from Table 2, even with attitude factored into the equation as the group variable, the results that emerge for the two prompts are very comparable. On Prompt 1, the negatively oriented group of raters emerged with a slightly higher mean score than the positively oriented group did in both OSM and PBM mediums. All four scores were in a narrow range of 11.19–11.52, close to the OSM/PBM figures presented in Table 1. No significance emerged on the *t* test where attitude was the grouping variable.

Similar results were obtained for Prompt 2. The range of scores was again quite comparable with the OSM/PBM figures in Table 1, with a score range of 13.31–13.78 obtained. In contrast with Prompt 1, the marking medium figures for Prompt 2 showed slightly higher scores being awarded by the positively oriented group, although the difference was small, and again with no significance recorded. We can thus conclude that no differences emerged between the two groups.

Table 2: Rater Attitude With Prompt Difficulty and Marking Method

Prompt	Marking method	Rater attitude	*N*	*M* (max. 24)	*SD*	*t* test results
1. Working in the fashion industry	OSM	negative	428	11.43	5.19	*ns*
		positive	204	11.22	5.82	
	PBM	negative	428	11.52	5.25	*ns*
		positive	204	11.19	5.64	
2. Clever or beautiful	OSM	negative	1,171	13.31	5.14	*ns*
		positive	597	13.64	5.29	
	PBM	negative	1,171	13.39	5.12	*ns*
		positive	597	13.78	5.47	

Correlations Between 2007 HKCEE English Language Papers

In this section, correlations between the different papers are presented for the two groups of raters. Table 3 (reproduced from Coniam, 2009a) presents the data from the 2007 HKCEE with regard to the correlation of Paper 1B2 (the paper under examination in the current study) with the other components of the HKCEE English language papers. The benchmark for correlations that the HKEAA aims for is the 0.8 level, since it is only at this level or above that correlations can be described as "strong" (Hatch & Lazaraton, 1991, p. 441). As can be seen, with the exception of the Speaking paper, all correlations are at the 0.8 level.

To give as full a picture as possible, correlations with different components of the 2007 HKCEE are now presented. Partial correlations were first conducted to examine the effect of attitude as a variable. A very small non-significant correlation of 0.021 emerged, indicating that rater attitude to the rating medium had very little effect.

Table 3: Correlations Between 2007 HKCEE English Language Papers

Correlation of HKCEE Writing Paper Task 2 with …	Correlation
2007 Paper 1A (Reading)	0.80
2007 Paper 2 (Listening & Integrated Skills)	0.81
2007 Paper 3 (Speaking)	0.72
2007 Paper 4 (School-Based Assessment [Oral])	0.83
2007 Subject Mark	0.90

Table 4 presents the results of the bivariate Pearson correlations. As can be seen from the table, all correlations were significant at the 1% level, with the majority above the 0.8 level. The lowest set of correlations achieved were between the two sections of the Writing paper — that is, Task 2 (the extended writing task which has been the major focus of investigation in the current set of studies) and Task 1 (the guided writing task). The correlations for the two groups were nonetheless very similar with a correlation of 0.76 for the negative raters and 0.75 for the positive raters. The correlation with the Reading paper, which the HKEAA takes as its anchor of reliability (King, 1994, p. 6), was 0.82 for the negative raters as against 0.83 for the positive raters.

Table 4: Correlations of OSM Task 2 With Other Components of the 2007 HKCEE English Language Papers for Negative and Positive Attitude Raters

	Writing task 1	Reading paper	Whole subject mark	PBM score
Negative attitude raters (N = 1,600)	0.76**	0.82**	0.85**	0.87**
Positive attitude raters (N = 800)	0.75**	0.83**	0.85**	0.87**

** $p < .01$

The "whole subject" mark comprises the composite score for the Reading paper, the Speaking paper, the Listening & Integrated Skills paper and the oral-oriented School-Based Assessment component but excludes the Writing paper to avoid skewing the results. The correlation of the OSM mark for Task 1 with the whole subject mark was a high 0.85 for both attitude groups. Finally, the correlation of OSM of Task 2 with PBM of the same task were also identical at a high 0.87. Once again, we find no significant differences between the two groups.

Rater Erraticness

In language performance tests (see, e.g., McNamara, 1996, p. 9) — with productive English language writing tests being considered weak

versions of such tests — the major statistical method of analysis accepted over the past decade has been multi-faceted Rasch measurement (MFRM), since it allows for situational factors such as rater severity, prompt difficulty and so on to be modeled and compensated for (Weir, 2005, p. 199). 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, the marking medium, and so on. The phenomena — the different situational factors — can be explicitly taken into consideration and modeled in constructing the overall measurement picture. In the current study, a five-faceted design was employed, modeling raters, test takers, input prompt materials, rating scales, and the marking medium. The computer program FACETS (Linacre, 1994) was used to perform the analysis. Table 5 presents the results for raters. The unit of measurement in Rasch analysis is the logit. These are measures which are centered at zero, with, in the current case, a measure of zero logits indicating a rater of average severity. Raters with positive logit scores are therefore severe while those with negative logit scores are lenient.

In Table 5, column 4 presents the infit mean square statistic, which describes model fit — "fit" essentially being the difference between expected and observed scores. Definitions of "fit" vary. "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, it can be seen that, with the exception of Raters 68 and 197, 22 of the 24 raters show good fit. Raters' logit values extend from +0.73 to –1.76, a range of 2.49 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. Some 3.42 logits was recorded in the Coniam (2008) study, with a 4.26 logit spread in Eckes (2005). The reliability of 0.99, however, indicates that raters are being reliably separated into different levels of severity.

Only two raters showed misfit. The worst fitting rater was Rater 68 who was a positive rater; Rater 197 was a negative rater. In Table 5, there is, however, no pattern to indicate that either group of raters showed an unequal degree of fit or that one group exhibited greater erraticness than the other.

Table 5: Raters' Measurement Report

Rater	Attitude	Logit value	Infit mean square	Model error
68	Positive	+0.47	1.42	.06
197	Negative	+0.32	1.34	.06
41	Negative	–0.13	1.22	.06
150	Positive	+0.72	1.19	.06
110	Positive	–1.08	1.18	.06
132	Negative	–1.34	1.10	.06
28	Positive	–1.76	1.08	.07
4	Negative	–1.25	1.07	.07
2	Positive	–0.65	1.02	.07
103	Negative	–0.33	1.01	.06
92	Negative	–0.41	0.97	.06
25	Positive	–1.12	0.94	.06
101	Positive	–0.39	0.93	.06
182	Negative	–0.22	0.91	.06
140	Negative	–0.85	0.90	.06
76	Negative	–0.28	0.89	.07
6	Negative	+0.32	0.88	.06
180	Negative	–0.31	0.88	.06
8	Positive	–0.31	0.84	.07
104	Negative	+0.73	0.82	.07
56	Negative	+0.22	0.78	.06
1	Negative	–0.08	0.78	.06
57	Negative	–0.75	0.78	.06
167	Negative	–0.42	0.75	.06
M		–0.37	0.99	.06
SD		+0.64	0.18	.00

Notes: RMSE: .06; Adj (True) *SD*: .65; Separation: 10.04; Reliability: .99;
Chi-square: 2320.8; *df*: 23; significance (probability): .00

While, in MFRM, rater severity is modeled along with the other facets, to give as full a picture as possible, rater severity figures for the two groups are presented in Table 6.

Table 6: Rater Severity Analysis for the Two Groups

Rater attitude	*N*	Logit value	*SD*	*t* test results
Negative	16	−0.30	0.55	*ns*
Positive	8	−0.51	0.82	
Both groups		−0.37	0.65	

As mentioned, raters extended in severity from the most severe rater at +0.73 to the most lenient at −1.76 logits. Given this range of 2.49 logits and a mean of −0.37 logits, the logit values of the negative and positive groups — which were also non-significant from *t* test results — at −0.30 and −0.51 logits can be seen to be quite comparable. This finding can therefore be interpreted as indicating that rater severity is not a significant factor between the two groups of raters.

Conclusion

The current study was framed in the context of a jurisdiction moving totally to the marking of public examinations from a paper-based mode to an onscreen mode. In the live 2007 HKCEE English Language Writing paper, all rating was performed on screen. The current study involved an investigation which compared the scores and performance of 16 raters with an essentially negative orientation toward the OSM process against 8 raters who exhibited a positive orientation.

Two hypotheses were investigated. The first hypothesis was that a negative attitude would impact upon marking in that test takers would receive lower grades from raters who viewed OSM negatively. By extending the analysis of previous studies (Coniam, 2009a, 2009b) and by factoring in *attitude* to the results of OSM versus PBM, results were returned that were comparable with those obtained from the previous study. When *t* tests were conducted using attitude as a grouping variable, no significant difference emerged between the two groups of raters with regard to either the prompt or the marking method. Further, correlations between the scripts rated by those with negative and positive attitudes were also very comparable with figures obtained from the 2007 examination as a whole. The first hypothesis was therefore rejected.

The second hypothesis was that raters who hold a negative attitude toward OSM would be more erratic than more positive-oriented raters.

As the results of the raters' MFRM output indicate, neither group showed any discernible pattern with regard to their fit data. There were two misfitting raters, with one — exhibiting the most misfit — being a positive attitude rater, while the second was a negative attitude rater. Consequently, the second hypothesis was also rejected.

Given that in 2012, all public examinations in Hong Kong will be marked solely on screen, it is important to ensure that the system is reliable. The results of the current study provide support that this is likely to be the case.

Note

1. The HKEAA Research and Development Committee is the executive committee that oversees the research activity of the HKEAA. Along with members from the HKEAA's governing council, the Committee comprises members from the local educational community such as school teachers, principals, and university professors.

Acknowledgment

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

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

Burns, R. B. (2000). *Introduction to research methods* (4th ed.). Frenchs Forest, NSW, Australia: Pearson Education Australia.

Cheng, M. Y. (1993). *Testing and re-testing in Hong Kong F.5 and F.6 English secondary classes*. Unpublished master's thesis, The University of Hong Kong, Hong Kong.

Coniam, D. (1991). Essay marking: A comparison of criterion-referenced and norm-referenced marking. *Institute of Language in Education Journal*, *7*, 154–164.

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. (2009a). A comparison of onscreen and paper-based marking in the Hong Kong public examination system. *Educational Research and Evaluation*, *15*(3), 243–263.

Coniam, D. (2009b). *Research into onscreen marking of English language composition writing: A study conducted by CUHK and HKEAA*. Hong Kong: Hong Kong Examinations and Assessment Authority.

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.

Fowles, D. E., & Adams, C. R. (2005, September). *How does assessment differ when e-marking replaces paper-based marking?* Paper presented at the 31st International Association for Educational Assessment conference "Assessment and the Future of Schooling and Learning," Abuja, Nigeria.

Geertz, C. (1973). Thick description: Toward an interpretative theory of culture. In C. Geertz, *The interpretation of cultures: Selected essays* (pp. 3–30). New York: Basic Books.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. Boston, MA: Heinle & Heinle.

Hong Kong Examinations and Assessment Authority. (2007). *2007 HKCEE English language: Examination report and question papers*. Hong Kong: Author.

King, R. (1994). Historical survey of English language testing in Hong Kong. In J. Boyle & P. Falvey (Eds.), *English language testing in Hong Kong* (pp. 3–29). Hong Kong: The Chinese University Press.

Legislative Council Finance Committee. (2005). *Grant to support the modernisation and development of the examination systems of the Hong Kong Examinations and Assessment Authority* (FCR(2005–06)33). Retrieved December 16, 2009, from http://www.legco.gov.hk/yr05-06/english/fc/fc/papers/f05-33e.pdf

Linacre, J. M. (1994). *FACETS: Rasch measurement computer program*. Chicago: MESA Press.

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

Powers, D., & Farnum, M. (1997). *Effects of mode of presentation on essay scores* (ETS Research Report RM-97-08). Princeton, NJ: Educational Testing Service.

Powers, D., Farnum, M., Grant, M., & Kubota, M. (1997). *A pilot test of online essay scoring* (ETS Research Report RM-97-07). Princeton, NJ: Educational Testing Service.

Powers, D., Kubota, M., Bentley, J., Farnum, M., Swartz, R., & Willard, A. E. (1998). *Qualifying essay readers for an online scoring network (OSN)*

(ETS Research Report RR-98-20). Princeton, NJ: Educational Testing Service.

Raikes, N., Greatorex, J., & Shaw, S. (2004, June). *From paper to screen: Some issues on the way*. Paper presented at the 30th International Association for Educational Assessment conference "Assessment in the Service of Learning," Philadelphia, U.S. Retrieved November 17, 2009, from http://www.cambridgeassessment.org.uk/ca/digitalAssets/113972_From_Paper_to_screen._Some_Issues_on_the_Way.pdf

Sturman, L., & Kispal, A. (2003, October). *To e or not to e? A comparison of electronic marking and paper-based marking*. Paper presented at the 29th International Association for Educational Assessment conference "Assessment in the Service of Learning," Manchester, U.K. Retrieved June 20, 2007, from http://www.aqa.org.uk/support/iaea/papers.html

Twing, J., Nichols, P. D., & Harrison, I. (2003, October). *The comparability of paper-based and image-based marking of a high-stakes, large-scale writing assessment*. Paper presented at the 29th International Association for Educational Assessment conference "Assessment in the Service of Learning," Manchester, U.K. Retrieved June 20, 2007, from http://www.aqa.org.uk/support/iaea/papers.html

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

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke, U.K.: Palgrave Macmillan.

Whetton, C., & Newton, P. (2002, September). *An evaluation of on-line marking*. Paper presented at the 28th International Association for Educational Assessment conference "Reforming Education Assessment to Meet Changing Needs," Hong Kong SAR, China.

Zhang, Y., Powers, D. E., Wright, W., & Morgan, R. (2003). *Applying the online scoring network (OSN) to advanced placement program (AP) tests* (ETS Research Report RR-03-12). Princeton, NJ: Educational Testing Service.

Appendix: 2007 HKCEE English Language Writing Paper, Task 2

(Reproduced with the permission of the Hong Kong Examinations and Assessment Authority)

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.

Essay Competition!

Win 6 weeks' work experience in the fashion industry.

Would you like to work

1. with a famous fashion designer;
2. on a popular fashion magazine;

OR

3. 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.