An Investigation into the Effect of Raw Scores in Determining Grades in a Public Examination of Writing

David Coniam
The Chinese University of Hong Kong

This article examines the effect on the grades assigned to test takers either directly through the use of raters' raw scores, or through the use of measures obtained through multifaceted Rasch measurement (MFRM). Using data from the Hong Kong 2005 public examination of writing, the current study examines how test takers' grades differ by comparing the results of grades from "lenient" raters against those of "severe" raters on the two systems for assigning grades–raw band scores and MFRM-derived scores. Examination of the results of a pair of raters indicates that the use of raw scores may produce widely different results from those obtained via MFRM, with test takers potentially disadvantaged by being rated by a severe rather than a lenient rater. In the Hong Kong English language public examination system from 2007 onwards, band scales are to be used extensively, as indeed they already are in many Asian countries. The article therefore concludes with a call for consideration to be given to how test takers' final grades may be derived from raw scores.

本研究は香港における公的試験のライティング・テストの採点に関する実証研究である。採点者の得点をそのまま使った場合と、多相ラッシュ・モデリング（ＭＦＲＭ）の得点を使った場合、成績の上でどのような違いがあるのかを調査したものである。香港で2005年度に実施された試験をデータとして使った。分析の結果、採点者の得点をそのまま使った場合には、より厳しい採点者によって受験者が不利を蒙る傾向があることがわかった。採点者の得点を使って最終成績をつける場合にはどうすればよいのかを論じて結論とした。

This article examines the use of raw scores obtained from the writing test of a public examination for Year 11 (the eleventh grade of schooling) test takers in Hong Kong. The current study draws on the methodology of Coniam (2005), who addressed an issue discussed by Weir (2005) on the notion of score validity, concerning the use of raw scores being an imperfect measure of test taker ability. Weir states "if FACETS is not being used in the evaluation of writing tests, I would want to know why not!" In the Coniam study, rater grade differentials on an oral test were investigated using novice raters. The current study extends the scope of the findings through data from a live Hong Kong public examination of writing using experienced raters.

With one major exception, rating scales are not a feature of English language public examination assessment in Hong Kong.[1] In the writing and oral public examinations, test takers are assessed using holistic, norm-referenced scales. As of 2007, the examination system in Hong Kong is, however, undergoing drastic changes in the English language elements (SCOLAR, 2003). This will involve the adoption of a standards-referenced, rather than a norm-referenced, approach to assessment, with scales and descriptors being used to rate test taker performance in English language examinations. In light of these changes to the Year 11 public examination, the purpose of the current study is to investigate how test takers' final grades differ depending on whether raw scores or Rasch-derived measures (Rasch, 1960) are used.

Raters and Raw Scores

In Hong Kong English language examinations, test takers' final grades 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 tests taken by the test takers, essentially the result is the raw score. The accuracy of the information obtained from raw scores has long been questioned, with the problems associated with their use having been discussed by a number of researchers. McNamara (1996, p. 122) presents a cogent discussion of some of the problems associated with the use of raw scores. Referring to studies by Linacre (1989) and Diederich, French, and Carlton (1961), he illustrates the variability in raw scores awarded to test takers, clearly stating that raw scores are "an unreliable guide to ability" (p. 118), and citing various reasons for this. He attributes, for example, variability in raters' assessment to a range of causes: rater (mis)interpretation of the rating scales and descriptors, rater freshness (or tiredness), and interpersonal factors where raters respond positively or negatively (albeit unintentionally) to certain gender, race, or personality types. Research conducted by Hamp-Lyons (1989) suggests that raters respond to cultural differences in writing, which is in part attributable to their own cultural and experiential background. Vann, Lorenz, and Meyer (1991) relate raters' responses to their gender as well as their academic discipline. Vaughan (1991) illustrates how raters' reactions to different language features may result in essays being awarded different grades.

Linacre (1989) suggests that the above-mentioned issues (which may affect test taker performance) are *facets*, which can–or indeed should–be taken into account, and be modelled when assessing test takers in performance tests. This is especially the case with the latter type of test, where many more factors need to be considered. With fixed-response test items–for which a limited set of answers are possible–there are likely to be few extraneous factors to be taken account of.

Major changes to the system by which writing scripts are rated in Hong Kong Year 11 public examinations are imminent–one crucial change involving the move to using rating scales. Given this, using data from the 2005 Hong Kong Certificate of Education (HKCE) examination, the current study sets out to examine the use of raw scores in a writing test and how test takers' grades compare when rated by a severe as opposed to a lenient rater. The current study involves a comparison of the use of raw scores with scores derived from statistical procedures such as multifaceted Rasch measurement (where situational factors such as prompt difficulty or rater severity may be modelled and compensated for; see below) when calculating test takers' final grades.

To restate, the hypothesis being addressed in the current study is therefore that the use of raw scores may substantially disadvantage test takers who are rated by severe rather than lenient raters, with those test takers receiving lower final grades–a situation which in some examination situations may result in failure rather than success on a test.

The Hong Kong School and Examination System

Hong Kong's model of education, although currently undergoing substantial revision, is modelled on the British system. There are 6 years of primary school, and secondary school operates on a 5+2 model with students being banded, or streamed, on entry to secondary school. There are three broad bands of ability, with each band covering approximately 33% of the student ability range.

Hong Kong's major public examination is the Hong Kong Certificate of Education (HKCE) examination, administered by the Hong Kong Examinations and Assessment Authority (HKEAA) at the end of Secondary 5 (Grade 11). In 2005, the candidature for English language was 82,078 (Hong Kong Examinations and Assessment Authority [HKEAA], 2005). There are four papers in the English language HKCE–Writing; Reading; Oral; and Integrated Reading, Writing, and Listening. The HKCE Writing paper–the focus of the current study–offers test takers three prompts. They have to select one and are allowed 70 minutes in which to write in the region of 300 words. Overall grades awarded on the HKCE English language paper are A to C (credit), D and E (pass), F and U (Fail). Grade C and above are the crucial grades since the University of Cambridge accepts these as a GCSE level pass. [2]

Figure 1 (from the 2005 HKCE examination) presents the prompt around which discussion in the current article centres.

> The Leisure and Cultural Services Department is planning to hold an international pop music festival in an open area very close to where you live. It has invited local residents to write letters expressing their views on the proposal.
>
> Write a letter to the Department giving your opinion and explaining the benefits and/or problems of holding the festival. If you wish, you may refer to one or more of the following in your letter:
>
> - noise levels
> - entertainment value
> - tourism
> - large crowds
> - hygiene and waste disposal
> - possible performers
> - opportunities for local musicians
>
> Begin your letter, "Dear Officer, ... " and sign it "R. Lee."

Figure 1. 2005 HKCE Writing Paper, Prompt 2

In the HKCE Writing paper, two raters assess each script independently with scripts currently pattern-marked on a single norm-referenced

9-point scale, with raters having to adhere to a specified pattern in terms of how many scripts can be allocated to a given point on the scale. Having to conform to a pattern mitigates, to an extent, the issue of severity since there are only so many high or low grades a rater may award. This changed, however, in 2007 when rating scales were adopted and raters were not constrained to a pattern.

Research Design

This section describes the data which made up the study and the methods used to analyse the data.

Data

The data used in the study were drawn from the live HKCE 2005 English language examination. Subsequent to the administration of the examination, 900 scripts (i.e., 300 scripts for each of the three prompts) were identified and extracted on the basis of the following three principles. First, that scripts should be drawn from the batches of markers with good statistics (i.e., good interrater reliability and a high correlation with other HKCE English language papers). Second, that scripts awarded the same grade by both markers should be selected since there would then be no differences between raters' raw scores. Third, that scripts selected should form a representative cross-section of ability across the whole candidature.

Nine markers then re-marked the three sets of 300 scripts. These were Hong Kong English teachers who had served as raters for the HKCE Writing paper for a number of years and had consistently achieved good rating statistics.

To prepare for the rating sessions, raters were first trained and their ratings standardised. After having familiarised themselves thoroughly with the new scales and descriptors, raters attended a training session where they rated a number of sample scripts illustrating different aspects of the scales and descriptors and a spread of ability across the 6-point scale. Because each script was double marked in line with standard HKEAA practice, 1,800 ratings were obtained for the current study; each rater assessed 200 scripts.

The subscales and descriptors used were those developed for the 2007 HKCE Writing Test (Note 2). The four subscales were:

1. Relevance and adequacy of content for purpose;

2. Accuracy and appropriacy of punctuation, vocabulary, language patterns;
3. Planning and organisation; and
4. Appropriacy of tone, style, and register; appropriacy of features for genre.

The subscales each had six levels, ranging from 1 (indicating the least able) to 6 (indicating the most able). For the subscales and descriptors, see HKEAA, 2007, pp. 104-105.

Methodology

As mentioned, the methodology in the current study involves a comparison of two composite scores. One of these was the rater's average of the four raw subscale scores. The second was obtained through multifaceted Rasch measurement (MFRM), a brief description of which will now be presented.

In classical measurement theory (CMT), test results cannot really be directly compared with one another. Consider for example, two Year 11 ESL classes. Last year's class scored 47% on their final exam; this year's class 43% on their (different) final exam. How are the two classes' scores to be compared? Is this year's class less able than last year's? Were the questions more difficult this year? Were the markers more severe in their judgements this year? We are not really in a position to answer any of these questions. Additionally, in CMT, test takers' results are not evenly spaced–despite the use of an apparently linear scale such as the percentage scale. Scores in the middle range are bunched together, while scores at the top and bottom end of the scale are disproportionately spread out (see Bond & Fox, 2007, pp. 24-26, for a cogent elaboration).

The use of the Rasch model enables all of these issues 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 result of using a Rasch model of measurement provides, in principle, independence from situational features (test takers, for example) in a particular test.

Consequently, results can be interpreted with a more general meaning. To return to the example in the above paragraph, test scores for different groups–such as last year's and this year's ESL classes–can be directly compared via Rasch measurement as the use of Rasch locates them on a single linear scale.

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, test taker ability, and rater severity levels. The phenomena or different situational factors can be explicitly taken into consideration and modelled in constructing the overall measurement picture.

While the focus in this study is on the rater, rater behaviour was examined in the context of the overall picture whereby it formed part of a three-faceted model of analysis, (i.e., raters, test takers, and prompts). The data presented in the paper is taken from the scores generated by the multifaceted Rasch analysis computer program FACETS (Linacre, 1994). In addition to logit measures, FACETS provides a "Fair Average" (see Linacre, 1997, p. 550, for details). The Fair Average is a more easily interpretable statistic in that logit values are converted back to the original rating scale, the 6-point scale in our case, rendering the output more easily interpretable by end-users. Because they are presented in the format of the original rating scale scores, the Fair Averages can be directly compared with the raters' original raw scores. Such a comparison forms the cornerstone of the methodology in the current study.

For an accessible overview of MFRM, the manner in which it may be conducted, and its results interpreted, the reader is referred to Bond and Fox (2007).

Results and Discussion

The analysis in this section centers on an examination of the differences between test taker scores using the two different methods of arriving at a final score, (i.e., the average raw score of the four subscales compared with the FACETS-provided Fair Average). For illustrative purposes, results will be presented for one pair of raters only: the pair of raters who showed the widest degree of divergence in terms of severity levels.

First, however, Table 1 presents the results derived through MFRM for all the raters. In this Table, Column 5 presents the infit mean square statistic, which describes model fit, "fit" essentially being the difference between expected and observed scores. "Perfect fit" is defined as 1.0, with an acceptable upper limit of fit stated as 1.3 (see Bond & Fox, 2007, pp. 285-286 for a discussion of limits of fit).

Table 1. Raters' Results

Rater	Logit values	Fair Average	Model error	Infit mean square	Notes	
181	**+2.19**	**2.53**	**0.06**	**0.98**	most severe rater	paired with rater 142
123	+0.59	3.02	0.06	1.10		
106	+0.46	3.06	0.06	0.80		
171	+0.40	3.08	0.06	0.90		
183	-0.09	3.23	0.06	0.94		
110	-0.35	3.32	0.06	0.97		
142	**-0.95**	**3.52**	**0.06**	**1.03**		paired with rater 181
153	-1.03	3.55	0.06	0.90		
188	-1.23	3.62	0.06	1.13	most lenient rater	
Mean	0.00	3.22	0.06	0.97		
S.D.	1.01	0.32	0.00	0.10		

Note. RMSE .06 Adj (True) S.D. 1.01 Separation 16.18 Reliability 1.00
Fixed (all same) chi-square: 2409.9 d.f.: 8 significance (probability): .00

As infit mean square values in Table 1 indicate, all nine raters were well within acceptable degrees of fit. Looking at Column 2 (logit values), raters show quite a spread of severity, extending from the most severe rater at +2.19 logits (Rater 181), to the most lenient (Rater 188) at -1.23. The range of rater severity is consequently wide. Taking 3 standard errors as a delineator of a "statistically distinct" level (see Wright & Masters, 1982, p. 92), the Separation index of 16.18 indicates that raters are being separated into distinct levels of severity. The raters' values in Column 2 are presented in logits. The Fair Averages are presented in Column 3 with logit values converted back to the original 6-point scale.

An analysis of the data will now be presented with regard to test takers' average band score compared with their Fair Average. The analysis presented centers on the pair of raters with the widest severity differential who rated the same test takers. These were Rater 181, with a measure of +2.19, and her partner Rater 142, with a measure of -0.95 logits. From the Fair Average scores in Table 1–where a lower score indicates a more severe rating–Rater 181 (whose Fair Average score is 2.53) can be seen to be one whole level more severe than Rater 142 (whose Fair Average score is 3.52). Their results are in bold type in Table 1 above.

It should be noted that Raters 181 and 142 co-rated 65 scripts: 22 test takers on Prompt 1, 27 on Prompt 2, and 16 on Prompt 3. The results and trends that emerge from the data hold good across all three prompts. To avoid overwhelming the reader with detail, however, only the largest data set (i.e., Prompt 2) is presented.

In Table 2 below, Column 2 provides the Fair Average. Two columns of data are then presented for each rater. The first column for each rater contains the rater's average raw band score from the four rating subscales; the second column presents the difference between the average raw band score and the Fair Average. A positive figure in a rater's second column indicates that the test taker would have received a higher (i.e., more lenient) score from that rater. A negative figure indicates a lower score, emerging from a more severe rating.

As can be seen from Table 2, the raters' tendency to severity or leniency is confirmed in the results that test takers would have received. Comparing the average raw scores against the Fair Averages, it can be seen that with Rater 181, all (100%) of her test takers would have received a lower grade. In contrast, only one (4%) of Rater 142's test takers would have received a lower grade, whereas 26 (96%) would have received a higher grade.

I would now like to explore further the extent to which the variation apparent in Table 2 above might be significant in determining a test taker's score on a test as a whole. As mentioned (in Note 1), band scales are only used on one test in Hong Kong–the English language teachers' Language Proficiency Assessment of Teachers (LPAT). On the test components that comprise the LPAT, test takers must reach level 3 of the 5-point scale on every scale, although they may still be awarded a pass with a 2.5 on one scale (Government of the Hong Kong Special Administrative Re-

Table 2. Prompt 2, Paired Raters 181 and 142

Test taker	Fair Average (FA)	Rater 181 (tendency to severity)		Rater 142 (tendency to leniency)	
		Average raw band score	Average raw band score minus FA	Average raw band score	Average raw band score minus FA
540	4.86	3.25	-1.61	6.00	+1.14
588	4.34	3.00	-1.34	5.25	+0.91
596	4.99	3.75	-1.24	5.75	+0.76
591	5.23	4.00	-1.23	6.00	+0.77
597	5.23	4.00	-1.23	6.00	+0.77
420	3.18	2.00	-1.18	4.00	+0.82
590	5.11	4.00	-1.11	5.75	+0.64
592	5.36	4.25	-1.11	6.00	+0.64
594	5.11	4.00	-1.11	5.75	+0.64
595	5.36	4.25	-1.11	6.00	+0.64
419	2.82	1.75	-1.07	3.50	+0.68
359	1.81	0.75	-1.06	2.50	+0.69
390	2.05	1.00	-1.05	2.75	+0.70
589	4.99	4.00	-0.99	5.50	+0.51
593	5.62	4.75	-0.87	6.00	+0.38
586	5.36	4.50	-0.86	5.75	+0.39
480	4.08	3.25	-0.83	4.50	+0.42
450	3.82	3.00	-0.82	4.25	+0.43
539	5.49	4.75	-0.74	5.75	+0.26
598	4.99	4.25	-0.74	5.25	+0.26
509	4.34	3.75	-0.59	4.50	+0.16
479	3.82	3.25	-0.57	4.00	+0.18
360	2.05	1.50	-0.55	2.25	+0.20
389	2.28	1.75	-0.53	2.50	+0.22
599	5.49	5.00	-0.49	5.50	+0.01
585	5.92	5.50	-0.42	6.00	+0.08
510	3.30	3.25	-0.05	3.00	-0.30

Note: (N=27)

gion, 2000). Obtaining two 2.5 level scores, or indeed any level 2 or lower score, results in an automatic failure grade being awarded on the LPAT. The writing test is regarded as one of the most demanding components of the LPAT (see Glenwright, 2002, for a discussion). The pass rate for the writing test is consistently one of the lowest across all five papers of the LPAT; in 2006, test takers achieved a proficiency attainment rate of 45.9% on this paper (HKEAA, 2006). Since the difference of half a band may therefore result in the difference between failing and passing, I would now like to explore this issue further.

Justification for half a band on the Hong Kong LPAT being taken as a determinant of "notable difference" can be seen to lie in the fact that the standard error of measurement (SEM) for the LPAT Writing Test is approximately 0.5 of a band (HKEAA, personal communication). This is comparable to the SEM for the IELTS Academic Writing module which, in 2005, was stated to be 0.37 of a band (IELTS, n.d.).

To underscore the importance of how half a band may stand as a "notable difference," Table 3 provides a summary of the differences between the Fair Averages produced by MFRM and those produced from the two raters' average raw band scores.

Table 3. Summary of Fair Averages Versus Raw Average Band Scores–Prompt 2

Leniency / Severity situation	Rater 181 (→ severity)	Rater 142 (→ leniency)
More lenient by half a band or more	0	14 (52%)
More lenient by less than half a band	0	12 (44%)
More severe by less than half a band	3 (11%)	1 (4%)
More severe by half a band or more	24 (89%)	0 (0%)
More lenient cases (total)	0 (0%)	26 (96%)
More severe cases (total)	27 (100%)	1 (4%)

(N=27)

As can be seen from Table 3, the two raters present almost a mirror image. If such results appeared on the LPAT, the consequences would be as follows. With half a band taken as criterial, 24 (89%) of Rater 181's test takers would have received a lower grade and potentially failed the LPAT whereas none of those rated by Rater 142 would have.

Conversely, for the test takers scored by Rater 142, 14 test takers (52%) would have been rated more than half a band higher, against none by Rater 181. If the half band score is crucial, over half of Rater 142's test takers might have been moved out of the potential failure zone, as against none of Rater 181's.

The implications of the differences between the two systems of rating are apparent: If a test taker were rated by a lenient rater such as Rater 142 as opposed to a severe rater such as Rater 181, the use of raw scores means that one test taker might "pass" the test while the other might well "fail." A discussion of this issue is provided by Coniam and Falvey (2001) in the context of the Hong Kong LPAT where simple raw scores are used to determine a final grade, and where a half-band score did, in certain cases, result in failure.

Conclusion

This study has examined the use of raw scores in the application of rating scales in the Hong Kong Certificate of Education (HKCE) 2005 Writing Test. The study has illustrated how the use of raw scores and measures derived through multifaceted Rasch measurement (MFRM) can produce markedly different results. The grades of two raters who assessed the same set of test takers were markedly different when the two methods of analysis were contrasted. Over half of the most lenient rater's test takers (52%) would have received a grade higher by half a band when this rater's raw scores were compared with MFRM-derived measures, with no test taker receiving a grade lower by half a band or more. In contrast, none of the most severe rater's test takers would have received a grade higher by half a band, although 89% would also have received a grade lower by half a band.

The current study has its limitations, however. The first of these lies in the fact that, to make its point, the study has been focusing on two extreme raters. An extension of the current study would possibly involve an examination of the "bigger picture" or how many test takers would have a different outcome (i.e., those who passed using raw scores, but failed using Rasch measures and vice versa) if MFRM had been used to determine students' proficiency. It is in such a situation that the effects of rater variance really become apparent–when fair ratings are not provided and when students' lives are unfairly affected. The study has also focused essentially on test takers who are affected by severe raters and who are receive a lower grade than they may merit. The converse is also true: that using raw scores rather than Rasch measures awards some test takers higher grades than they deserve. Nonetheless, because more anguish is caused by test takers who fail when they should pass rather than vice versa, the focus in the current study is what it is.

Further, the current study has drawn on data from public examinations. Practical applications lie in the use of Rasch measurement (underpinned by an understanding of Rasch principles, Rasch, 1960) in school-based situations. However, convincing English language teachers of the value of certain statistics and getting them to use them represents something of a challenge. Popham (2006) comments, for example, on the temptation "to characterize any sort of test-related topic as 'too technical' for teachers" (p. 25). Nonetheless, it is achievable. The Centre for Assessment & Development (CARD) at the Hong Kong Institute of Education (http://www.ied.edu.hk/card) has a project running with about 100 primary and secondary schools in Hong Kong, with the objective of raising teachers' awareness of assessment targets and how targets may relate incrementally to other targets at higher levels. The project draws strongly on Rasch measurement principles, both as a technical tool as well as one that is delegated down to the teacher level with teachers using Rasch measurement in their evaluation of the tests they produce as an indicator of student progress and achievement. Between 2006 and 2007, CARD ran a total of 28 workshops (for more than 1,600 teachers) on Rasch measurement principles with hands-on practice in using Winsteps. Follow-up feedback indicated that individual teachers, and even whole schools, began to experiment with Rasch-based school assessment initiatives (see http://www.ied.edu.hk/card).

Given the results discussed in the current study, as Hong Kong moves towards adopting the use of scales and descriptors in rating test takers in its English language examinations when a standards-referenced approach to assessment is adopted in 2007, this issue of raw scores and the consequent disparity of results through rater severity is one which merits serious consideration. Given the fact that HKCE Grades A to C are recognised as a GCSE level pass, although a lower grade may not result in failure (as it can do with the LPAT), the consequence of being rated by a severe rather than a lenient rater may make the difference between a test taker achieving a D rather than a C on the HKCE.

The use of band scales and descriptors are now the currently accepted method by which most speaking and writing tests are rated, with the practice being adopted in most countries across Asia. The implications from this small-scale Hong Kong study can therefore be extended beyond the Hong Kong context, and constitute an issue that needs to be considered by many educational and assessment bodies moving towards rating with scales and descriptors. The advantages of MFRM for analysing rating in speaking and writing tests are not new, as has been mentioned. The current study has attempted to underline the value of such a system in deriving test takers' results, suggesting that, if examination bodies adhere to a system whereby raw scores are the main determinant of a final grade, this may be doing test takers a disservice.

Acknowledgement

I would like to thank the Hong Kong Examinations and Assessment Authority for access to test takers' scripts from the 2005 HKCE English language examination.

David Coniam is a professor in the Faculty of Education at The Chinese University of Hong Kong, where he is a teacher educator, working with ESL teachers in Hong Kong secondary schools. His main publication and research interests are in language assessment, language teaching methodology, and computational and corpus linguistics.

Notes

[1] The one exception where scales and descriptors are currently used in a Hong Kong English language examination is the Language Proficiency Assessment of Teachers (LPAT) test for English language teachers (Government of the Hong Kong Special Administrative Region, 2000). The test consists of five papers. Of these, Speaking, Writing, and the Classroom Language Assessment test (a performance test conducted in a teacher's live classroom) are rated using scales and descriptors, with raw marks determining the final score. On the Speaking, Writing, and the Classroom Language Assessment test components, LPAT test takers must reach level 3 of the 5-point scale on every scale, although with a 2.5 on one scale one will still be awarded a pass (Government of the Hong Kong Special Ad-

ministrative Region, 2000). Obtaining any level 2 or lower score results in an automatic failure grade being awarded.

[2] Pass rates for the 2005 HKCE were: Grades A–C: 10.5%; Grades A–E: 70.7% (Available at: http://www.hkeaa.edu.hk/doc/fd/2005cee/ceexamstat05_1.pdf).

References

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

Coniam, D. (2005). Raw scores as examination results: How far can they be relied upon? Paper presented at the Association of Language Testers in Europe Second International Conference, Berlin, 19-21 May 2005.

Coniam, D., & Falvey, P. (2001). Awarding passes in the Language Proficiency Assessment of English Language Teachers: Different methods, varying outcomes. *Education Journal, 29* (2), 23-35.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factors in the judgment of writing quality*. Princeton, NJ: Educational Testing Service.

Glenwright, P. (2002). Language proficiency assessment for teachers: The effects of benchmarking on writing assessment in Hong Kong schools. *Assessing Writing*, *8* (2), 84-109.

Government of the Hong Kong Special Administrative Region. (2000). *Language benchmark assessment for teachers–English language: Syllabus specifications explanatory notes, specimen questions with suggested answers, scales and descriptors*. Hong Kong: Government Printer.

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

Hong Kong Examinations and Assessment Authority (HKEAA). (2005). *2005 HKCEE entry statistics.* Retrieved October 11, 2007, from http://www.hkeaa.edu.hk/doc/fd/2005cee/ceexamstat05_2.pdf

Hong Kong Examinations and Assessment Authority (HKEAA). (2006). *Language Proficiency Assessment for Teachers (English language) 2006: Assessment report.* Retrieved October 11, 2007, from http://eant01.hkeaa.edu.hk/hkea/redirector.asp?p_direction=body&p_clickurl=otherexam%5Fbycategory%2Easp

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

International English Language Testing Service (IELTS). (n.d.). *Test performance 2005.* Retrieved October 11, 2007, from http://www.ielts.org/teachersandresearchers/analysisoftestdata/article234.aspx

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago: MESA Press.

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

Linacre, J. M. (1997). Communicating examinee measures as expected ratings. *Rasch Measurement Transactions, 11*(1), 550-551. Retrieved October 11, 2007, from http://www.rasch.org/rmt/rmt111m.htm

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

Popham, W. J. (2006). *Mastering assessment: A self-service system for educators*. New York: Routledge.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. The Danish Institute of Educational Research, Copenhagen.

SCOLAR. (2003). *Action plan to raise language standards in Hong Kong*. Hong Kong: Government Printer.

Vann, R., Lorenz, F., & Meyer, D. (1991). Error gravity: Faculty responses to errors in the written discourse of nonnative speakers of English. In Hamp-Lyons, L. (Ed.), *Assessing second language writing in academic contexts* (pp. 181-195). Norwood, NJ: Ablex.

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.

Weir, C. (2005, May). *A socio-cognitive approach to test validation*. Plenary presented at the Association of Language Testers in Europe Second International Conference, Berlin.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press, Chicago.