40th IAEA 2014 Conference

Role of Assistant Examiners in Marker Behaviour Modification during Onscreen Marking

CHEUNG Kwai Mun Amy PhD
kmcheung@hkeaa.edu.hk
LO Yuen Kwan Wincy
yklo@hkeaa.edu.hk
Hong Kong Examinations and Assessment Authority

Abstract

This study investigated the effectiveness of the assistant examiners' (AEs) under instant feedback conditions during onscreen marking (OSM) of the written components of Chinese Language and English Language at Primary 6 (Grade 6) in the Hong Kong's Territory-wide System Assessment (TSA). The OSM was first adopted for marking TSA papers in 2008. It is held **centrally** for two weeks per year at a Hong Kong Examinations and Assessment Authority marking centre. During OSM, the AEs monitor the markers' performance by check-marking the marked scripts on the spot and identifying problematic scripts using these instant statistical figures: 1) marking speed; 2) percentage of third-mark triggering; 3) leniency/severity of each marker; 4) rating distribution of each marker. AEs also have meetings with all markers on standardization using the aforementioned figures to contextualize live scripts. AEs' interventions are immediately triggered when markers show inconsistency and/or idiosyncratic ratings. For this study, a questionnaire was completed by AEs on the use of OSM. Both the quantitative data from the instant statistics and the qualitative data from the questionnaires indicated that modification of marker behaviour was evident and a strong congruence was observed among markers over the past few years.

Key words: onscreen marking, rater characteristics, language assessment

1 Introduction

This study investigated the effectiveness of the assistant examiners (AEs) under instant feedback conditions during onscreen marking (OSM) of the written components of Chinese Language and English Language at Primary 6 (Grade 6) in Hong Kong's Territory-wide System Assessment[1] (TSA). TSA is held annually in June (end of academic year) and onscreen marking (OSM) of Primary 6 (Grade 6) TSA written papers is conducted **centrally** over a two-week period in July from 9am to 5pm in the assessment centres of the Hong Kong Examinations and Assessment Authority (HKEAA). Onscreen marking (OSM) has been in place since 2008. A Chief Examiner (CE) is appointed from the tertiary sector to take charge of Assistant Examiner (AE) and Marker training in conjunction with the HKEAA's Manager-in-charge of the subject level. (Although the OSM personnel are referred to as 'markers' in the administrative literature, they are in fact 'raters' in terms of assessment literature.) About 110 markers and 14 AEs for Chinese Language and around 60 Markers and

[1] Territory-wide System Assessment is a low-stakes survey of the performance of students at Primary 3 (Grade 3), Primary 6 (Grade 6) and Secondary 3 (Grade 9) levels in Chinese, English and Mathematics. The TSA aims to provide the Hong Kong Special Administrative Region (HKSAR) Government and school management with information on **school** standards in key learning areas.

eight AEs for English Language are recruited yearly. Any teacher wishing to serve as a Marker has to attend a four-hour training session which includes discussion of marking criteria[2] and supervised rating of language samples. After this, teachers review their ratings focussing on how well they describe the samples in question.

2 Role of Assistant Examiners in OSM

2.1 Rating Standardization Scripts

Before the OSM starts, AEs are assigned to rate a number of scripts for standardization. These scripts are used to ensure markers' consistency and marking quality at the three stages of the marking process[3]: 1) Training; 2) Qualification – to assess whether markers have met the set requirements before commencing marking; and 3) Control – to monitor markers' quality during marking; scripts are randomly assigned to each marker throughout the process.

Scripts to be used for standardization are drawn from a stratified random sample (N=300 from a total of some 500 participating schools with a total student population of 60,000). The following are examples of Primary 6 (P.6) English standardization scripts. These scripts are randomly rated by eight AEs with 'overlapping marking' in place. Each AE is only required to rate a maximum of 90 scripts in a three-hour session. In other words, AEs are only required to rate about one-third of the total number scripts for standardization. This arrangement is cost effective and helps lower the chance of rater fatigue reducing reliability. For each AE, his/her scripts overlap with one other AEs so that they form the unbroken chain of overlap required for later Rasch analysis. A similar arrangement is adopted for marking P.6 Chinese standardization scripts.

2.2 Verifying Ratings of Standardization Scripts

An expert panel including all AEs is assembled to verify the ratings of the standardization scripts. The other panel members include a Chief Examiner, the Manager-in-charge of the level and a subject officer of HKEAA. Rasch's 'fair average' (FA) ratings (Linacre, 1987-2013) are obtained from ratings by the AEs and verified by the expert panel. Adjustments to ratings are made based on members' professional judgment in cases where members do not agree with the Rasch FA. From the experience over the last few years, less than 5% of the FA ratings required adjustment. Adjustments were normally made in some scripts with FA ratings of 1.5 or 1.6 which had been rounded up to '2'. After being judged by the expert panel, the scripts in question were adjusted to a rating of '1'.

---

[2] Marking criteria for Chinese writing include: content, structure, sentence, vocabulary, punctuation and wrong words. Marking criteria for English writing include: content and language.

[3] These stages involve the use of standardization scripts and Rasch derived fair average ratings for the standardization scripts. This method of benchmarking has been proved valid and reliable (Cheung & Chang, 2009).

2.3 Maximizing Consistency among Markers

To maximize consistency among markers, four vital functions of AEs are performed as follows:

1) standardized training of markers
2) qualifying of markers to ensure they are within set tolerance when rating standardization papers
3) check-marking the markers throughout the entire OSM period
4) intervening markers who are found problematic in rating

3 Instant Feedback during OSM

Each live script is double marked (see Table 1 'OSM Interface'). During the entire process of OSM, assistant examiners (AEs) are required to checkmark markers' scripts for no less than 5% of the total scripts from each marker and identify problematic scripts using a number of instant statistical figures built in the computer system (see Table 2 'Marking Statistics'). The mechanism of check-marking and instant feedback ensures marking quality and maximizes marker congruence. It is important to note that neither instant figures nor AE feedback targeted particular scripts. The feedback was kept general to avoid creating fake reliability as an artefact of feedback. The main instant statistical figures[4] are as follows:

Table 1. OSM Interface

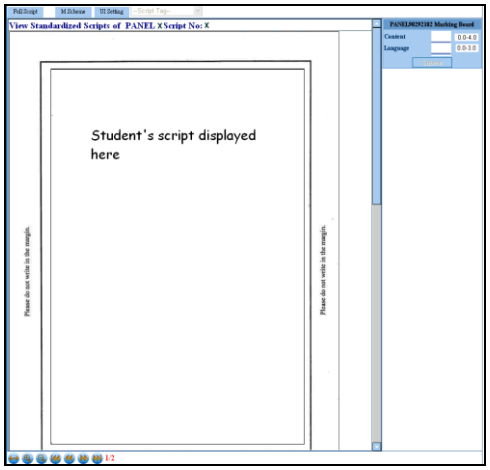
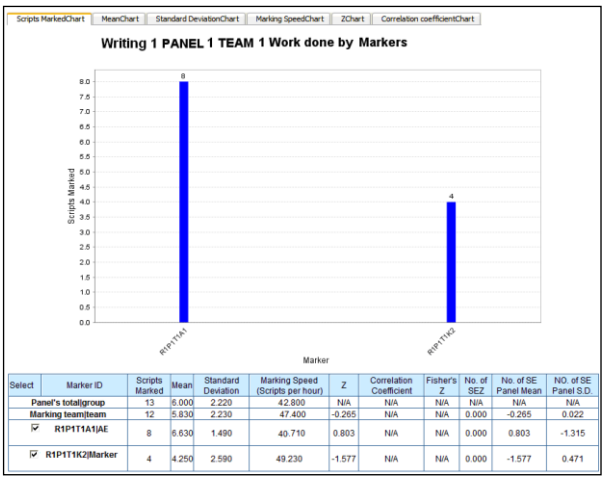


Table 2. Marking Statistics



Select	Marker ID	Scripts Marked	Mean	Standard Deviation	Marking Speed (Scripts per hour)	Z	Correlation Coefficient	Fisher's Z	No. of SEZ	No. of SE Panel Mean	NO. of SE Panel S.D.
	Panel's total|group	13	6.000	2.220	42.800	N/A	N/A	N/A	N/A	N/A	N/A
	Marking team|team	12	5.830	2.230	47.400	-0.265	N/A	N/A	0.000	-0.265	0.022
☑	R1P1T1A1|AE	8	6.630	1.490	40.710	0.803	N/A	N/A	0.000	0.803	-1.315
☑	R1P1T1K2|Marker	4	4.250	2.590	49.230	-1.577	N/A	N/A	0.000	-1.577	0.471

3.1 Figure (1): Marking speed

This figure shows the number of scripts marked per hour

3.2 Figure (2): Marker leniency/severity

Deviation of marker's mean from the mean of all markers (panel mean) in terms of no. of standard error (SE) of mean (how lenient or severe each marker was by comparing his/her mean against the mean for the entire panel of total markers); acceptable range is $-2 < x < +2$.

[4] To minimize the effects of measurement errors, these figures require at least 180 scripts rated by each marker.

If the figure is far < -2, this means a very harsh marker. If the figure is far > +2, this means a very lenient marker.

$$\text{No. of SE Panel Mean} = \frac{\text{Marker's Mean} - \text{Mean of All Markers}}{\text{SD of All Markers' Mean}}$$

3.3 Figure (3): Rating distribution of each marker

Deviation of marker's standard deviation (SD) from the SD of all markers (panel SD) in terms of no. of standard error of SD (the size of the SD of each marker was as compared to the SD of the panel); acceptable range is $-2 < y < +2$. If the figure is far < -2, this means the marker gives a narrow range of ratings while the figure is far > +2, this means the marker gives an excessive no. of ratings on opposite ends of the scale.

$$\text{No. of SE Panel SD} = \frac{\text{Marker's SD} - \text{SD of All Markers}}{\text{SD of All Markers' SD}}$$

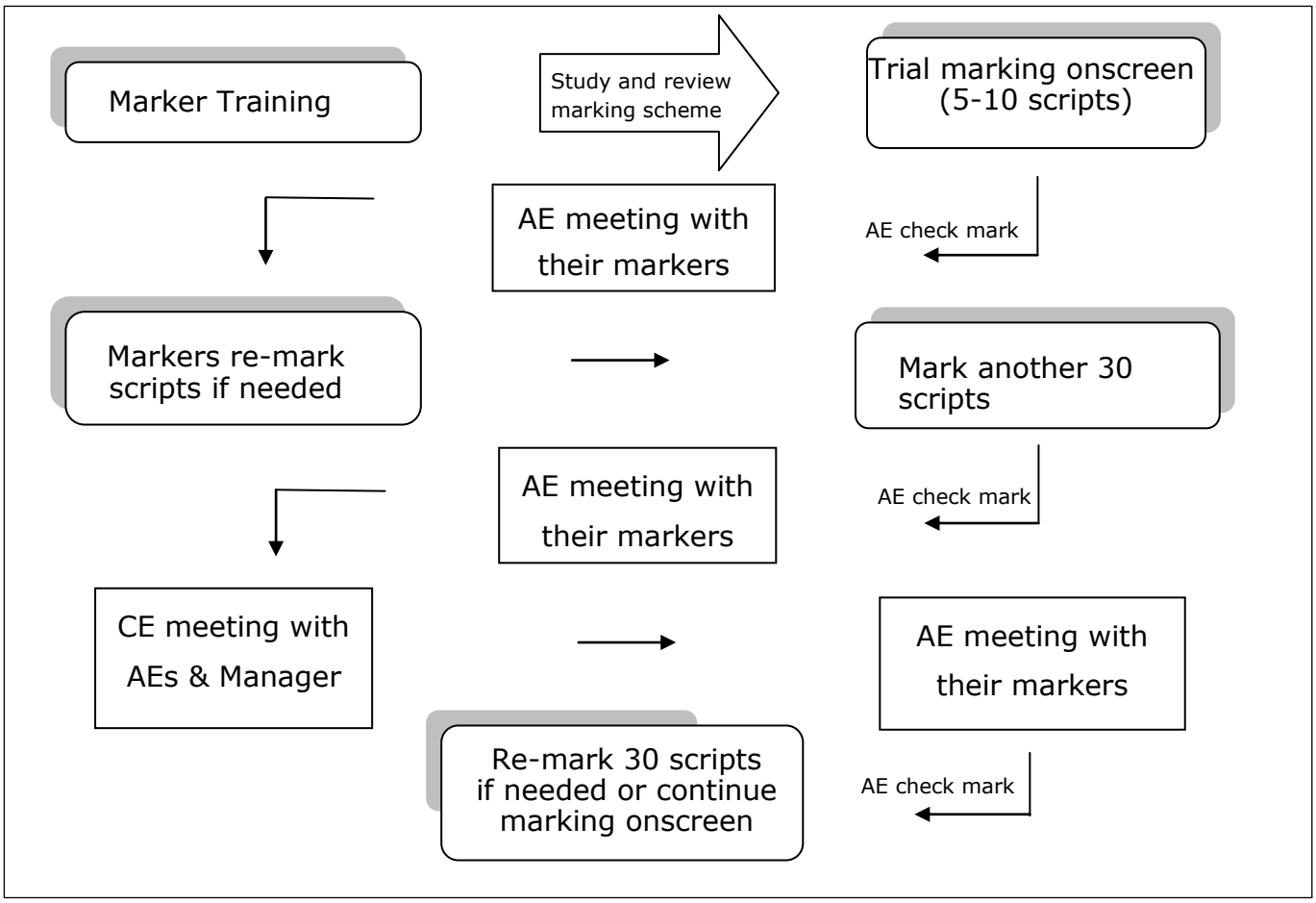
3.4 Figure (4): Percentage of third-mark triggering

When the discrepancy of the two assigned markers is beyond the acceptable range (the range decided by Manager) of major marking criteria (i.e. content), third marking will be triggered and carried out by the relevant AE. Most markers used the given statistical figures, i.e. marker severity and rating dispersion for reference but a few experienced markers used these figures improperly. Markers who did this adjusted their scores up or down based solely on their severity and rating distribution figures without referring back to the actual scripts. Such markers sometimes adjusted the wrong ratings, e.g. reducing the wrong 7 or boosting the wrong 2. This behaviour became evident in the high percentage of third-mark triggering due to these markers. Under these circumstances, these markers were identified and their AE scrutinised their scripts and intervention was in place where necessary.

4 Exercising Professional Judgment on Standardization during OSM

Throughout the entire OSM, not only the aforementioned instant figures are used for standardization. Professional judgments are also exercised to ensure marking quality. Panel group meetings with markers on standardization are essential since intervention by AEs can only be triggered statistically after a marker has rated more than 180 scripts. The purposes of the meetings are: 1) to collect data on common misjudgments and errors made by the markers, 2) to gather enquiries and information from markers on borderline cases and problematic scripts and 3) to disseminate decisions made (if any) at the meetings with CE during OSM.

Diagram 1. A flowchart detailing the processes for standardization among markers in the initial stages of the OSM



5 Findings – Quantitative

Basically, the four types of statistical figures in the OSM system have already ensured the students' ratings are within the acceptable range. However, one of the other purposes for this study is to analyze individual markers' performance in depth and use these analyses to improve markers' performance. As a result, the accuracy of rating students' writing has increased. The following findings were collected from the data from Primary 6 English Writing component of the previous TSA.

5.1 Case 1: A Lenient Marker Giving a Narrow Range of Ratings

Intervention was in place when the AEs found that their markers' statistical figures were out of the acceptable range. This could only be ascertained statistically after they had rated more than 180 scripts. However, some markers did not go out of range until much later. The purpose of intervention was to bring errant markers back into congruence. Normally, the markers had to review the scripts they had already rated before they could rate new scripts.

In case 1, a total of three interventions were required before Marker A began to rate reliably (see Table 3 for details). AE1 had her first intervention with Marker A (as at 5pm on Day 2). Marker A's ratings were a little lenient (No. of SE of Panel Mean = 2.195: a slightly higher proportion of ratings of 7 were given) and she gave a narrow range of ratings (No. of SE of Panel SD = -2.160: the marker gave a disproportionate high amount of 4s). AE1 scrutinised the scripts with possible problematic ratings (ratings of 4 and 7 in this case) and identified scripts which tended to induce problematic ratings using her professional judgment. She then suggested the scripts for which Marker A was required to review and amend the ratings. Adjustments were made and Marker A's statistics were back in acceptable ranges (as at 2pm on Day 3).

The second intervention took place at 9:17am on Day 4 when Marker A's statistical figure on No. of SE of Panel SD was far below -2 (i.e. -3.003), meaning Marker A had tended to narrow her range of ratings (where she gave a proportionally high amount of 6s). Adjustments were made when scripts with ratings of 6 were further reviewed.

The third intervention took place at 10:15am on Day 4 (No. of SE of Panel Mean = 0.464, No. of SE of Panel SD = -2.462). AE1 reminded Marker A that she had tended to cluster her ratings around the median but at the same time, a new batch of scripts was allocated to Marker A and she was reminded to stay within range. Eventually, Marker A's performance became stable and her ratings stayed within the acceptable ranges after the third intervention (No. of SE of Panel Mean = 0.758; No. of SE of Panel SD = -1.25). Supply of a new batch of scripts can serve to break a marker out of an undesirable habit (e.g. giving too many 4s). The new scripts also increase the size of the data set allowing rating patterns to become more visible.

Table 3. Statistical Figures of Marker A in Marking Panel 1 Scripts

Date	**Time**	**No. of SE Panel Mean**	**No. of SE Panel SD**	**Analysis of Ratings (0-7)**
Day 2	17:00	**2.195**	**-2.160**	A little lenient 2.195 (>+2): a slightly higher proportion of 7s A narrow range of ratings -2.160 (<-2): a disproportionally high amount of 4s
		1st Intervention		
Day 4	09:17	-0.670	**-3.003**	A narrow range of ratings -3.003 (<-2): a proportionally high amount of 6s
		2nd Intervention		
	10:15	0.464	**-2.462**	A narrow range of ratings -2.462 (<-2): tended to cluster her ratings around the median
		3rd Intervention		
	12:30	0.831	-2.014	Marker A's performance became stable and her ratings fell within the acceptable ranges
	15:20	0.758	-1.250	

Marker A had no problem in rating the scripts in Panel 2 (same writing question) from Day 5 to Day 9. The two statistical figures: No. of SE of Panel Mean (-1.38 to 0.791) and No. of SE of Panel of SD (0 to 0.919) were inside the acceptable ranges. Marker A's rating performance stabilized after she had rated one third of her total scripts.

5.2 Case 2: A Harsh Marker Giving Widely Dispersed Ratings

Only two interventions were required before Marker B rated reliably. (See Table 4 for details.) AE2 had his first intervention (as at 5pm on Day 2) with Marker B when the marker's ratings were very harsh (-3.266). AE2 scrutinised the scripts with possible problematic ratings (proportionally more ratings of 0 and less ratings of 6 in this case) and identified the problematic scripts using his professional judgment. He then suggested the scripts which Marker B was required to review and amend the ratings where necessary. Adjustments were made and Marker B's statistics on severity became normal.

The second intervention took place at 2:45pm on Day 3 when Marker B's statistical figure on No. of SE of Panel SD was far above +2 (i.e. +3.198), meaning Marker B had a tendency to give widely dispersed ratings (she gave an excessive number of ratings on opposite ends of the scale). Adjustments were made when scripts with ratings of 0 and 7 were first reviewed, followed by scripts with other ratings (i.e. ratings of 2 and 6, and ratings of 3 and 5). Marker B's statistics were approaching acceptable ranges (as at 5pm on Day 3). After the second intervention (No. of SE of Panel Mean = -0.258; No. of SE of Panel SD = 2.010), Marker B's ratings began to fall within the acceptable ranges.

Table 4. Statistical figures of Marker B in marking Panel 1 scripts

Date	Time	No. of SE Panel Mean	No. of SE Panel SD	Analysis of Ratings (0-7)
Day 2	17:00	**-3.266**	-1.188	Severe -3.266 (<-2): proportionally more ratings of 0 and less ratings of 6
		1st Intervention		
Day 3	14:45	-0.754	**3.198**	Widely dispersed ratings +3.198 (>+2): excessive number of ratings on opposite ends of the scale
		2nd Intervention		
	17:00	-0.258	2.010	Marker B's performance became stable and his ratings fell within the acceptable ranges
Day 4	10:20	-0.133	1.878	

Marker B had no problem in marking the scripts in Panel 2 (same writing question) from Day 5 to Day 9. His two statistical figures: No. of SE of Panel Mean (0.916 to 1.865) and No. of SE of Panel of SD (-1.554 to 0) were within the acceptable ranges. As with the case of Marker A, Marker B had to rate one third of his total scripts in order to get himself familiarized with the marking scheme so that his marking performance became steady.

5.3 Effects of AE Intervention

The experience with Marker A and Marker B indicates that timely intervention by AEs serves to modify marker behaviour in the sense of making their ratings more valid and reliable. It can cure both inconsistency and over-consistency. However, there is a lot of variation in the amount of intervention required to stabilize a marker's rating behaviour into a valid and reliable pattern.

6 Findings – Qualitative

To add a descriptive dimension to the quantitative findings mentioned in Section 5, qualitative data were collected from all the 22 AEs of Primary 6 Writing components in Chinese Language and English Language of the previous TSA. The following is a summary of their views on the role of AE during the OSM.

6.1 Professional Feedback from AE

As part of their duties, AEs were required to check-mark and provide timely feedback where necessary. Some markers were already veteran language teachers with ample experience in marking TSA. It should not have been an easy task for the AEs to convince these markers to make amendments to their ratings. However, the OSM statistical figures served as objective evidence, offering a strong support to back up the AEs when they made judgments about marker performance. They trusted that the system could 'standardize the ratings of all markers' and helped markers 'review the marked scripts and make adjustment easily.'

- A number of AEs stated that: The statistical figures were very concrete and reliable in tracking the overall picture of markers' performance. They showed when the markers were lenient or severe or when they tended to give too many median ratings. AEs could identify markers' problems instantly, e.g. some markers had problems giving certain ratings such as 2 and 3.
- Most AEs also said that they had relied heavily on the figures. They used the figures from time to time to give feedback and keep track the markers' performance.

6.2 Marking Quality

The statistical figures played an important role in maintaining the marking quality. When asked about ways to ensure their marking quality, (apart from frequent re-visiting of marking schemes and exemplars), all AEs revealed that they had made reference to the statistical figures (Mean and SD) to check severity and rating distribution.

- A few AEs stated that: OSM provided concrete figures on which to evaluate markers' performance. Reference to figures was also the most tactful way to request markers to re-mark their scripts. Some had to re-mark scripts which were marked in a particular timeslot, e.g. Monday afternoon, or some had to re-mark scripts with ratings of two. This occurred when and wherever markers were found problematic.
- Some AEs mentioned that: They could retrieve some scripts where markers had made errors repeatedly. They found that this evidence was very convincing to support their point of view and thus persuade markers to re-mark their scripts.

6.3 Effect on Teachers' Professionalism

- Some AEs pointed out that self-reflection was occurring throughout the entire OSM process: While check-marking, they were able to view the ratings given by other personnel (i.e. markers, the other AE from the same group and the CE) on each script. They then compared these with their own ratings and evaluated their own judgment against the set standards. In doing so, they helped monitor their own marking and check-marking quality. If they found any inconsistencies, they would consult the manager-in-charge or the CE. This mechanism could serve for monitoring any marking personnel, not just markers.
- A few veteran AEs also commented that: Their duty was not confined to ensuring marking quality but also involved helping develop teachers' judgment of students' work. This was especially true for those who marked idiosyncratically.

6.4 Pros and Cons of OSM

Generally, all AEs indicated that OSM was an effective and efficient way of ensuring marking quality.

- Some AEs pointed out that: OSM was very efficient; as compared to paper-based marking, they stated that OSM allowed second marking to happen within a very short period of time. This mechanism enabled them to monitor marking quality and identify problematic markers and problematic scripts immediately because third marking would be triggered instantly if their ratings were found inconsistent with other markers.
- A few AEs added that: It was a better and easier way to perform AE's duties.
- Some AEs' comments show that the OSM ensures the data integrity of the TSA and they stated that: No live scripts were found missing.
- A number of AEs added that: Check-marking was faster and more comprehensive. They could review the scripts easily and the statistical figures could show the markers' performance instantly.
- An AE added that: There was a short message sending (SMS) system in the OSM and it was very useful to handle markers who were passive. These markers might feel embarrassed and uncomfortable when they were given verbal comments or they were told that their marking was sub-standard. Hence, using SMS avoided causing embarrassment.

All AEs believed that OSM helped markers keep on track and avoid erratic marking. When asked about their preferred mode of marking, all markers favoured OSM. For the limitations of the OSM, nearly all of them found that their eyes got tired very easily.

7 Conclusion

Before the implementation of OSM (i.e. when marking was paper-based), AEs took a long time to identify problem markers. Hence, the feedback given to the markers in question was slow compared to the OSM situation. Pre-OSM AEs were required to check-mark the markers and provide professional support when they found the markers were not marking according to the marking schemes, e.g. the markers were too lenient, severe and/or erratic. This involved check-marking a substantial number of scripts in order to see the patterns of

markers; however, OSM enabled accomplishment of the same goal with a smaller volume of check-marking because statistical data was instantly available for reference. The design of OSM system with instant statistical data has already ensured that ratings given were within the acceptable range. Moreover, with AE's timely feedback to markers using the instant statistical figures, the system has helped raise the performance of many markers, from good to excellent.

Both the quantitative data from the instant feedback system and the qualitative data from the AEs indicated that using OSM, the difficulties markers encountered could be resolved faster and problematic markers could be identified early enough to minimize their effects on reliability and validity. The AEs also held that the OSM instant feedback system facilitated self-reflection by markers and AEs while marking or check-marking. Hence, it did modify the marking behaviour of both markers and AEs. The AEs commented that a strong congruence had been observed over the last few years. According to the reports given by the AEs since 2010, about 85% of markers received Grade 'B'[5] or above as rated by their AEs.

In assessing students' writing skills, markers are essential, since it is their judgment which actualises the rating scale in terms of showing how good a performance is. (Although the OSM personnel are referred to as 'markers' in the administrative literature, they are in fact 'raters' in terms of assessment literature.) Rater characteristics are of particular interest as far as the nature and extent of variability in performance assessment is concerned because rating scales by their very nature require raters for their implementation. In future, studies on rater characteristics under the instant feedback system during the OSM are worth exploring, e.g. rater fatigue, selective attention to one or more characteristics of the performance being rated (such as grammatical accuracy, choice of words, sequence of events, causal relationship), the effect of 'jarring' errors on markers, such as 'I go to shopping' (I go shopping), 'the party is very well' (the party is great) and 'I very like' (I like it very much) (Cheung, 2010).

References

Cheung, K. M. (2010). *Reliability and validity in practice: Hong Kong's Key Stage 3 oral assessment.* Macquarie University, Australia, Unpublished PhD thesis.

Cheung, K. M. & Chang, R. (2009, September). *Investigating reliability and validity in rating scripts for standardization purposes in onscreen marking.* Presented at the 35th International Association for Educational Assessment (IAEA) Conference: Brisbane, Australia.

Linacre, J. M. (1987-2013). *FACETS: Many-facet Rasch Measurement computer program. Version 3.71.3*. Chicago, IL.

[5] Performance descriptors of Grade 'B'

- Rated consistently and most of the control scripts are rated correctly
- Understood the marking criteria and followed the marking schemes most of the time
- Communicated effectively with AE and other marking personnel most of the time