Using on-screen marking (OSM) data parameters to gauge markers' marking performance

Louis Yim

The 44th International Association for Educational Assessment (IAEA) Annual Conference, 9 – 14th September 2018, Oxford, UK

Contents

- Rationale behind the study
- A brief introduction to the SEAB's OSM flow
- Aims of the research
- Research methodology
- Quantitative findings
- Qualitative findings
- Conclusions

Rationale behind the study

- In many on-screen marking (OSM) systems, quality checks on markers, e.g. random check and monitoring scripts, are an indispensable part to monitor markers' marking performance;
- As the percentage of quality checks within the total no. of live scripts is usually low, these spot checks might not be able to characterise markers' performance in full;
- Secondly, the marking standard of senior markers who perform checks on junior markers might not be in line with that of the Chief Marker (CM) even after standardisation, which further conceals markers' intrinsic marking performance;
- The main thrust of this research is to make use of the CM's reference standards and the data collected via the OSM to gauge all markers' marking performance based on several OSM data parameters;
- Such *de facto* evidence of marking performance, e.g. the statistics of intrinsic markers' capabilities, should help refine approaches/strategies of markers' deployment for any future on-screen marking.

A brief introduction to the SEAB's OSM flow

Paper based exam

Script collection

Scripts being scanned

Scanned scripts are being marked on-screen at marking centre

- ✓ Standardisation;
- ✓ Live marking – random check + monitoring scripts;
- ✓ Closing of marking project;
- ✓ Data download from the OSM platform for analyses, if needed

Aims of the research

This study aims to conduct an on-screen marking research under the usual OSM conditions, and investigate several OSM parameters acquired from a *post-marking* data analysis to gauge markers' marking/checking performance with respect to the reference standards garnered from the Chief Marker. Markers' OSM experience will be captured at the end of the study. The purpose is to help refine approaches/strategies of markers' deployment for any future on-screen marking.

Research methodology

- 90 scripts were anonymised and prepared for the study, with each script comprises the source-based case study and structured essay question genres;
- The Chief Marker (CM) on-screen marked the 90 scripts to establish the reference standards prior to the main research;
- 5 markers (3 markers + 2 snr. markers) were invited to a 4-day research exercise, who went through the OSM training and standardisation;
- During live marking, each of the 3 markers was assigned the same set of 90 scripts to mark; the 2 snr. marker then carried out a random check on markers' scripts for the entire exercise;
- Monitoring scripts by item were distributed constantly during live marking to monitor markers' marking performance;
- On completion of the exercise, all 5 markers were asked to complete a questionnaire to feedback on their OSM experience. This was followed by a focus group discussion.

Quantitative findings

Analyses and results

Analyses objectives:

- Investigate the findings on markers' general marking accuracy via the inferential statistics of marks between markers' 1st marking attempts and the reference standards for all item scripts;
- Investigate the OSM parameters of:
  - ✓ markers' marking speed;
  - ✓ marking accuracy (via rmse) of monitoring scripts and live research scripts;
  - ✓ markers' strengths and weaknesses (via rmse) of different questions' genres;
- Compare all rmse related parameters against the reference standards garnered from the CM's marking of the same set of scripts;
- Discuss a scatter plot of marking accuracy vs. marking speed

Note: All analyses were processed via a suite of in-house programs written in Stata® software version 14.

Analyses and results – Inferential statistics based on markers' first marking attempts

- A paired sample t-test between markers' first marking attempts and the reference standards:

Variable	Pair-wise Scenario	$r_{A/B}$	t-value (p)	Interpretation of p for the pair
Snr. Marker 01's mark	Reference mark vs. OSM mark	0.974	-0.800 ($0.10 < p \leq 1$)	No evidence against H_0
Snr. Marker 02's mark	Reference mark vs. OSM mark	0.895	-0.520 ($0.10 < p \leq 1$)	No evidence against H_0
Marker 03's mark	Reference mark vs. OSM mark	0.784	0.00 ($0.10 < p \leq 1$)	No evidence against H_0
Marker 04's mark	Reference mark vs. OSM mark	0.779	-0.430 ($0.10 < p \leq 1$)	No evidence against H_0
Marker 05's mark	Reference mark vs. OSM mark	0.761	0.00 ($0.10 < p \leq 1$)	No evidence against H_0

- ✓ The means and s.d. (not shown) between the reference and OSM marks are very close to each other for all markers;
- ✓ $r_{A/B}$ are generally high, with SMK01 and SMK02 being very high;
- ✓ All markers' marks are not statistically significantly different from the reference marks;
- ✓ In other words, all markers' marking standards are generally in line with the reference standards

Analyses and results – OSM marking speed

- Definition of marker's marking speed:

$$Marking\ speed = \frac{total\ no.\ of\ item\ scripts}{duration\ in\ hour}$$

- The unit is *item script per hour* instead of *script per hour;*
- Marking speed expressed in rank order for both markers and senior markers:

Role	Marking/checking speed's rank order [1 = fastest; 5 = slowest]
Snr. Marker 01	1
Snr. Marker 02	5
Marker 03	4
Marker 04	2
Marker 05	3

Note: Only live scripts are included in the marking speed calculation.

- Speed's dynamic range = 24 to 31 *item script/hr* ;
- Snr. Marker 02 is the slowest amongst all markers, perhaps he/she is more thorough with the checking.

Background to markers' performance based on the rmse

$$\text{Root Mean Square Error (rmse)} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}$$

- The square root of the average differences between the predicted values ($\hat{y}_j$) and observed values (y_j);
- its value ranges between 0 and ∞;
- indifferent to the direction of errors;
- the lower the value, the better the accuracy;
- is sensitive to outliers;
- large errors from markers have a disproportionately large effect on the rmse value.

Analyses and results – rmse of monitoring scripts

- The notion of using rmse is to characterise individual markers' marking accuracy based on the *de facto* evidence collected empirically from the *in situ* marking;
- Findings of the monitoring scripts' rmse expressed in rank order. The calculation of rmse is based on marks awarded to the monitoring scripts compared with those from the reference standards:

Role	rmse of monitoring scripts' rank order (1 = most accurate; 3 = least accurate)
Marker 03	1
Marker 04	2
Marker 05	3

Note: rmse's dynamic range = 0 to 0.7

- The rank order shows that Marker 03 was the most accurate, and Marker 05 is the least accurate when marking monitoring scripts;
- The usefulness of markers' rank order is more apparent when more markers are in the marking panel.

Analyses and results – rmse of live research scripts

Role	rmse of live research scripts' rank order (1 = most accurate; 5 = least accurate)
Snr. Marker 01	1
Snr. Marker 02	2
Marker 03	4
Marker 04	3
Marker 05	5

Note: rmse's dynamic range = 0 to 1.20

- Two tiers of rmse performance are apparent from the findings, i.e. senior markers' rmse rank order is at the top 2, and markers' rmse rank order is at the bottom 3;
- There is a reverse in rank order between Marker 03 and Marker 04 when compared to that in the monitoring scripts.

Analyses and results – Markers' strengths & weaknesses of different questions' genres

- A dissection of markers' accuracy with respect to different question genres based on the rmse (c.f. reference standards):

Q. Genre / Role	rmse of live research scripts' rank order (1 = most accurate; least accurate)
Source-based case study	
Snr. Marker 01	1
Snr. Marker 02	2
Marker 03	3
Marker 04	4
Marker 05	5
Structured essay	
Snr. Marker 01	1
Snr. Marker 02	4
Marker 03	5
Marker 04	2
Marker 05	3

Note: rmse's dynamic range = 0 to 1.20

- ✓ Amongst markers, the rank order for *Source-based case study* is the same as that of the monitoring scripts;
- ✓ The entire markers' rank order for *Structured essay* is very different from the *Source-based genre*, with Marker 04 being the 2nd most accurate and Marker 03 being the least;
- ✓ The results suggest that markers do have their strengths and weaknesses when it comes to marking different question genres

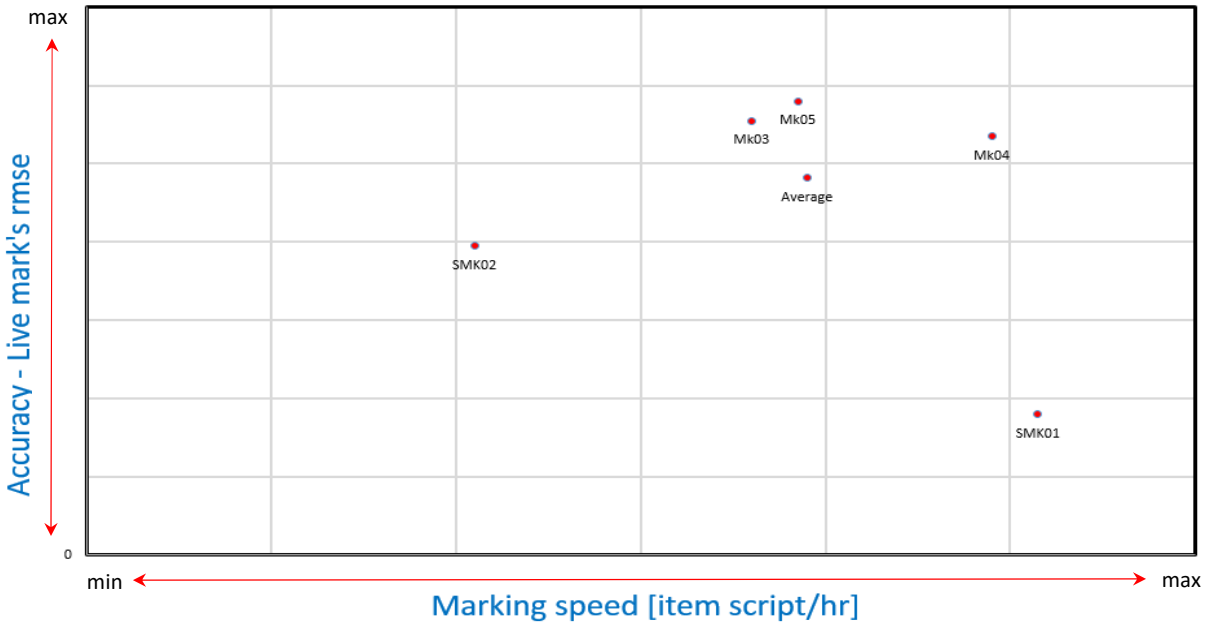
Analyses and results – Rank order dashboard for different OSM parameters

Role	Marking speed's rank order [1 = fastest; 5 = slowest]	Monitoring scripts' rmse [1 = most accurate; 3 = least accurate]	Live research scripts' rmse [1 = most accurate; 5 = least accurate]	Source-based case study genre's rmse [1 = most accurate; 5 = least accurate]	Structured essay genre's rmse [1 = most accurate; 5 = least accurate]
Snr. Marker 01	1	-	1	1	1
Snr. Marker 02	5	-	2	2	4
Marker 03	4	1	4	3	5
Marker 04	2	2	3	4	2
Marker 05	3	3	5	5	3

An overview of markers' performance of different OSM parameters expressed in rank order.

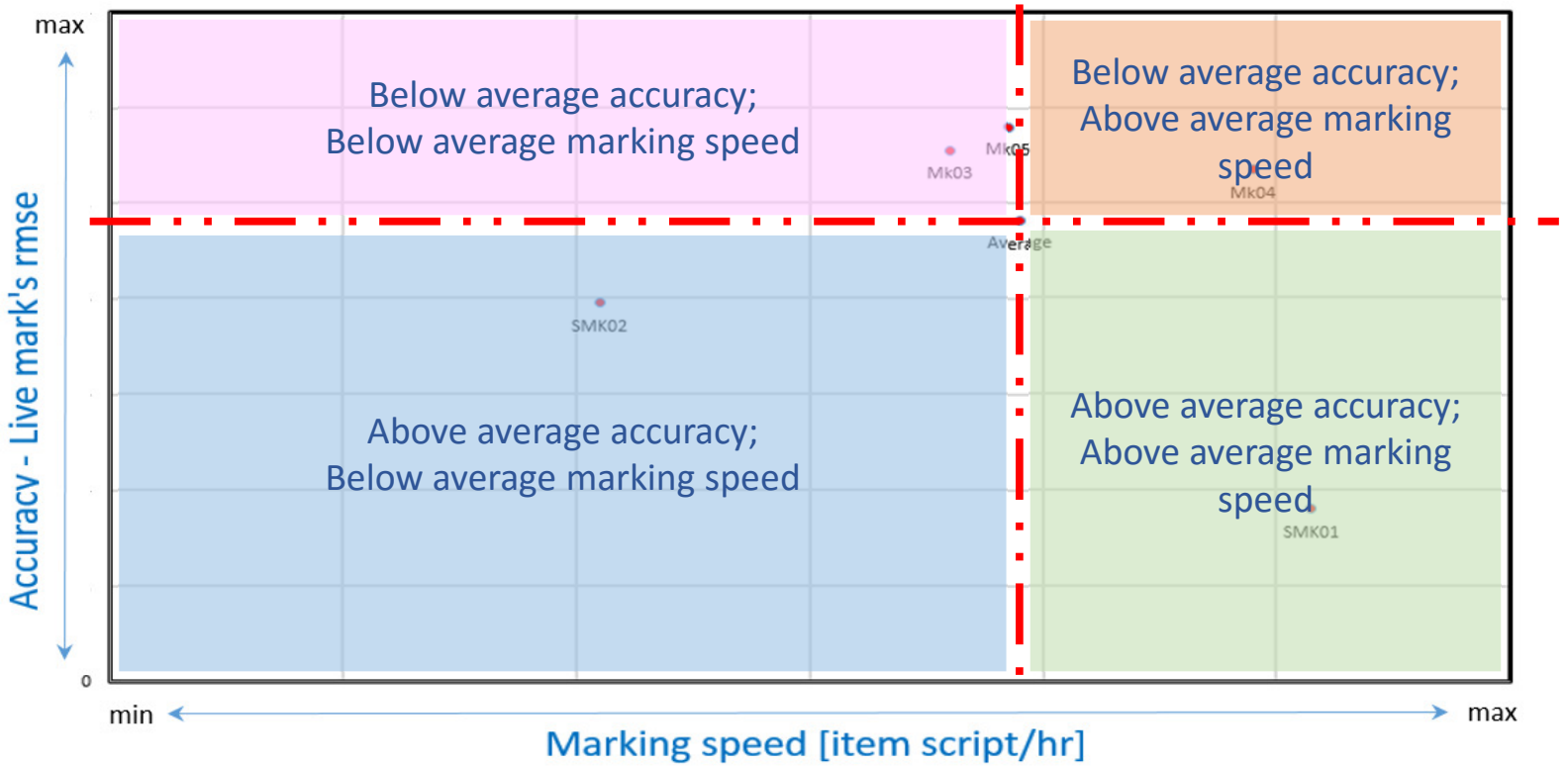
Analyses and results – Scatter plot of marking speed vs. overall markers' rmse

- From time to time, markers' disparate performance in different aspects paints a mixed picture of individual markers' capabilities;
- A systematic way of amalgamating such information to represent markers' performance two-dimensionally could be via a scatter plot of *marking accuracy* versus *marking speed*.
- A scatter plot of *Accuracy - Live mark's rmse* vs. *Marking Speed* is as follows:



Analyses and results – Scatter plot of marking speed vs. overall markers' rmse (2)

- Scatter plot of *Accuracy - Live mark's rmse* vs. *Marking/Checking Speed* with regions of competencies:



- Initially, the 'average' rmse and speed are used to establish the regions of competencies. With more data/information over time, an appropriate rmse and speed will be used to establish the regions accordingly or even set those values to help raise the markers' performance.

Qualitative findings

Markers' feedback on their OSM experience

- All markers and senior markers rated positively on their overall OSM experience;
- They felt the system helps reduce the amount of administrative and logistical work, e.g. adding marks together and counting scripts. It also helps prevent the loss of scripts as they are stored and retrieved digitally;
- They also felt that the system is easy to use and they have enough technological skills to resume to the normal marking pattern, albeit being new;
- Senior markers felt that more in-depth hands-on training would help improve their familiarisation with more complicated tasks during standardisation and quality checking processes;
- Findings on the pre- and post-marking surveys suggested that all markers and senior markers reported a higher preference rating for OSM compared to their prior expectation.
- All participants articulated that both mental challenge and exhaustion (both physical and mental) were expected when it comes to marking. Using the OSM system to mark did not add to extra challenges and exhaustion c.f. its paper-based counterpart.

Limitations of the study

- Under live marking condition, the number of scripts marked by markers would be much more than just 90 full scripts, the factor of fatigue over many days of marking may not be reflected genuinely in this research exercise;
- All rmse and marking speed results were only reported in ordinal level instead of interval/ratio level for presentation purpose. To assess markers' performance under normal circumstances, both ordinal and interval/ratio levels should be assessed to give a fuller picture of the markers' performance.

Conclusions

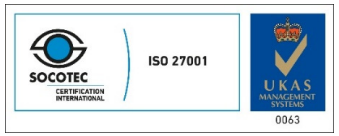
- Based on the *post-marking analyses* of several OSM parameters, this research helped dissect the markers' marking performance with respect to the reference standards;
- The information could be used to feedback to markers such that training could be provided in time to consolidate their less developed area(s). Alternatively, for more immediate tasks, markers could be deployed to mark question genre(s) at which they are strong to increase the marking accuracy;
- Such *de facto* evidence collected on markers' marking performance could help refine approaches/strategies of markers' deployment for future on-screen marking;
- More hands-on training would be given to new senior markers on standardisation and quality checking processes to expedite their familiarisation of the system;
- Markers generally rated positively on the OSM system, and found it rather user friendly even though using it for the first time. They felt the system helped reduce the amount of administrative and logistical work.

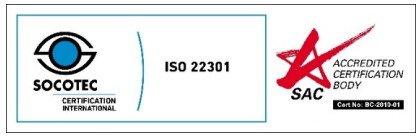
Bibliography

- Cheung, K.M.A. and Lo, Y.K.W.: Role of Assistant Examiners in Marker Behaviour Modification during Onscreen Marking (OSM), *2014 IAEA Conference Proceedings*, Singapore.
- Drasgrow, F., Luecht, R., and Bennett, R.: Technology and testing. Educational Measurement, ed. Brennan, R. Westport (CT): ACE/Praeger, 471 – 515.
- Drave, N.: Marker 'fatigue' and marking reliability in Hong Kong's Language Proficiency Assessment for Teachers of English (LPATE), *2011 IAEA Conference Proceedings*, Philippines.
- Haggie, D.: Onscreen marking: introduction strategy and examiner response, *32nd IAEA Conference Proceedings*, Singapore, 2006.
- Hopkin, R., Johnson, M., Shiell, H., Bell, J. F. and Raikes, N.: Extended essay marking on screen: Is examiner marking accuracy influenced by marking mode?, *36th IAEA Conference Proceedings*, Bangkok, 2010.
- Hudson, G.: Giving candidates a fairer deal in examinations and tests through electronic marking, *32nd IAEA Conference Proceedings*, Singapore, 2006.
- Hudson, G.: A quality control framework for electronic marking, *36th IAEA Conference Proceedings*, Bangkok, 2010.
- Myers, K.: Developments in On-screen Assessment Design for Examinations, *32nd IAEA Conference Proceedings*, Singapore, 2006.
- Shaw, S.: Essay Marking on-screen: implications for assessment validity, e-Learning, Vol. 5, Number 3, 2008. Available online: http:// www.wwwords.co.uk/ELEA

Singapore Examinations and Assessment Board

A trusted authority in examinations and assessment, recognised locally and internationally





Acknowledgement

The author would like to thank members of the research team for their contributions to the research exercise. Mohammad Faizal Zainal, Sharifah Mufidah Mohamed Aljunied, Yeo Yen Peng, Ng Xue Ting, Noelle Ho, Wee Tian Lu, Ng Siow Chin, Tan Hwa Mei.