A report on the effectiveness of double marking in the HKDSE Chinese Language writing examination

2019

Introduction

1. To address public concerns about rater effects and to ensure the quality and reliability of marking, the Authority adopts double marking for items that require markers' judgment. As double marking requires a considerable amount of resources, it is critical to ensure the efforts are worthwhile. This study aims at finding out whether double marking is significantly more reliable than single marking in the 2018 HKDSE Chinese Language writing exam.

2. Selecting more accurate markers is another way of improving marking efficiency. In this study, we also investigated whether a marker's marking experience and previous marking performance are good indicators of his/her marking accuracy.

The double marking scheme for the HKDSE Chinese Language writing exam

3. Each answer script is marked by two markers separately. If the marks assigned by the two markers differ within 11 marks, the final mark is the average of the two marks. If the difference exceeds 11 marks, the script will be marked by a third marker and the final mark is the average of the closest, highest two. If the difference among the three marks still exceeds 11 marks, the script will be marked by a fourth marker, who is an Assistant Examiner and the final mark is the average of the closest, highest two.

4. In order to ensure the marking quality, some answer scripts from each marker are selected for check marking. Check marking is conducted by more experienced markers, and their marks are considered to be more reliable and are used as a reference of marking accuracy. Check markers include Chief Examiners, Assistant Examiners, subject managers, and some experienced markers.

5. In the On-screen-marking (OSM) system II, there are two programs of assigning answer scripts to check markers. First, for each marker in each marking session, 3 scripts (each from one of the high, middle, and low mark ranges) are selected for

check marking. In addition, if a large mark discrepancy is found between two markers, the script will be check marked.

The data set

6. The 2018 HKDSE Chinese Language writing exam data was used in this study. The data set consisted of 7195 cases. Each case had three different scores. "M1" denotes the mark given by the first marker; "C1" denotes the mark given by the check marker which is considered as the reference in this study; and "FM" is the final mark and end result of the double marking scheme. The minimum score a candidate may receive is 0, and the maximum score is 103.

The research question and methodology

7. To answer the question whether double marking is more reliable than single marking, we need to compare double marking and single marking results with student's "true score". After discussion with subject experts, check mark is considered as our best estimate of the students' true score. So we compared M1 and FM with C1 in this study.

8. Besides descriptive statistics and correlations among M1, FM and C1, two linear regression analyses were performed to examine to how similar M1 and FM are to C1. The first regression equation is shown below:

$$C1 = a_1 + b_1 M1$$

The closer the regression coefficient b_1 is to 1, and the closer the intercept parameter a_1 is to 0, the more similar M1 is to C1. R2 is a coefficient of determination in regression analysis. It is the proportion of the variance in the dependent variable that is replicated by the independent variable(s). The closer R2 is to 1, the more variation in C1 can be explained by M1.

The second regression equation is similar to the first one, except for the independent variable is changed from M1 to FM:

$$C1 = a_2 + b_2\ FM$$

9. To improve the efficiency of the current marking scheme, selecting more accurate markers may reduce the need for a third and fourth marking. Two possible indicators of marking accuracy were investigated: a marker's years of marking experience and his/her previous marking performance.

10. With regard to marking experience, markers were classified into 2 groups, more-experienced and less-experienced. As for markers' previous performance, it was first rated by Assistant Examiners on a scale of 1 to 5, where 5 represents good performance and 1 represents poor performance. A marker was then classified as having good performance if he or she was rated 4 or 5 and did not get any 1 or 2 in the past 3 years. Otherwise, the marker was classified as having an average performance.

11. A regression analysis was conducted to identify whether these two variables are good indicators to identify more accurate markers. The regression equation is shown below:

$$C1M1_{ABSDIFF} = a + bEXP + cPER$$

"$C1M1_{ABSDIFF}$" denotes the absolute difference between C1 and M1. "EXP" denotes markers' marking experience (1 = more experience and 0 = less experience) and "PER" denotes markers' previous performance (1 = good performance and 0 = average performance). The larger the absolute value of the coefficients *b* and *c*, the better the variables can indicate makers' accuracy.

Results

12. The descriptive statistics are shown in Table 1. M1 contributed to the final mark in 2966 cases (41%), where the discrepancy between M1 and the second marker is smaller than 11 marks. The distribution of M1, FM and C1are fairly similar, with FM slightly closer to C1. For cases where M1 did not contribute to the final mark (59% of the sampled candidates), the distribution of FM is more similar to C1, indicating that double marking may be more accurate than single marking for these candidates.

Table 1 Descriptive statistics of the data set

Descriptive		N	Mean	Std. Deviation	Minimum	Maximum
M1 contributed to the final marks	M1	2966 (41%)	55.86	10.44	0	87
	C1		55.49	10.17	3	87
	FM		55.77	10.22	0	87
M1 did **not** contribute to the final marks	M1	4229 (59%)	55.17	11.22	9	90
	C1		55.95	9.80	0	89
	FM		56.10	9.81	0	100
All cases	M1	7195 (100%)	55.45	10.91	0	90
	C1		55.76	9.96	0	89
	FM		55.97	9.98	0	100

13. Table 2 shows the correlation between the three score variables M1, C1, and FM. As one may expect, when both C1 and M1 contributed to the final marks, both M1 and FM had an almost perfect correlation with C1 (>0.99) and the correlation between M1 and C1 was also extremely strong (0.975).

14. In the case when M1 contributed to the final marks while C1 did not, the correlation between M1 and C1 (0.950) was slightly lower than that between FM and C1 (0.966). Discrepancy of M1 from C1 was observed.

15. For the case when M1 did not contribute to the final marks, the correlation between M1 and C1 was only 0.662, while the correlation between FM and C1 remained high at 0.976. This implied that marks of the first marker might in many cases differ from the “true score”.

16. For all the cases, the correlation pattern was essentially an average of the two aforementioned scenarios.

Table 2 Correlation between three score variables in four scenarios

	N	M1 and C1	FM and C1	M1 and FM
Both C1 and M1 contributed to the final marks	2309 (32%)	0.975	0.993	0.994
M1 contributed to the final marks	2966 (41%)	0.950	0.966	0.994
M1 did **not** contribute to the final marks	4229 (59%)	0.662	0.976	0.623
All cases	7195 (100%)	0.776	0.972	0.771

17. In addition, regression models provide further details about how much double marking and single marking results deviate from the reference "true score". Tables 3a and 3b show the results of the two regression analyses respectively.

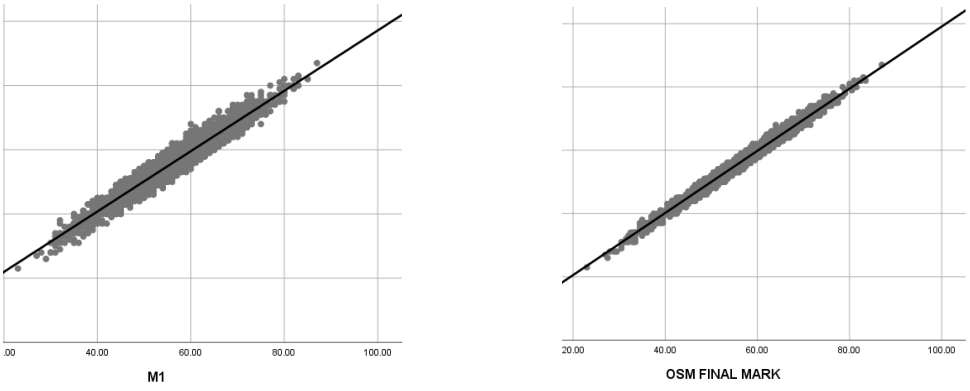
Table 3a Summary of the linear regression analysis: C1 = a_1+b_1 M1

	N	R^2	a_1	b_1
Both C1 and M1 contributed to the final marks	2309 (32%)	0.950	3.17	0.940
M1 contributed to the final marks	2966 (41%)	0.903	3.78	0.926
M1 did **not** contribute to the final marks	4229 (59%)	0.438	24.04	0.578
All cases	7195 (100%)	0.602	16.48	0.708

Table 3b Summary of the linear regression analysis: C1 = a_2+b_2 FM

	N	R^2	a_2	b_2
Both C1 and M1 contributed to the final marks	2309 (32%)	0.987	0.927	0.982
M1 contributed to the final marks	2966 (41%)	0.934	1.846	0.962
M1 did **not** contribute to the final marks	4229 (59%)	0.952	1.261	0.975
All cases	7195 (100%)	0.944	1.511	0.969

18. In cases where both C1 and M1 made contribution to the final mark, as shown in Table 3a, M1 is very close to C1, with a R^2 of 0.950, an intercept of about 3, and a slope of 0.94. But when compared with Table 3b, FM is even closer to C1, with a R^2 of 0.987, an intercept of only 0.927, and a slope of almost 1. This implies that a second marker did help make final marks closer to the "true scores". As shown in Figures 1a and 1b, data points tended to cluster along the regression lines in both figures but in Figure 1b, the spread of data points was extremely small.

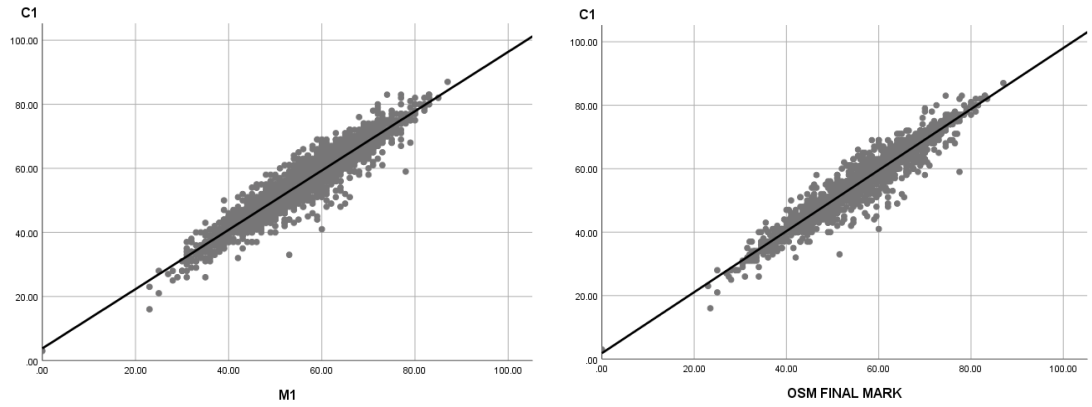


(a) (b)

Figures 1a and 1b Regression lines of (a) M1 predicting C1 and (b) FM predicting C1 when both C1 and M1 contributed to the final marks

19. When only M1 made contribution to the final marks, M1 could still to a large extent predict C1, with a R^2 of 0.903, an intercept of almost 4, and a slope of 0.926. The high R^2 value, small intercept value and a slope close to 1, altogether implies that M1 marks were in many cases very similar to C1 marks. Again, FM did an even better job in resembling the "true scores", with a higher R^2 of 0.934, a smaller intercept of 1.846, and a slope of 0.962. As shown in Figures 2a and 2b, the spread

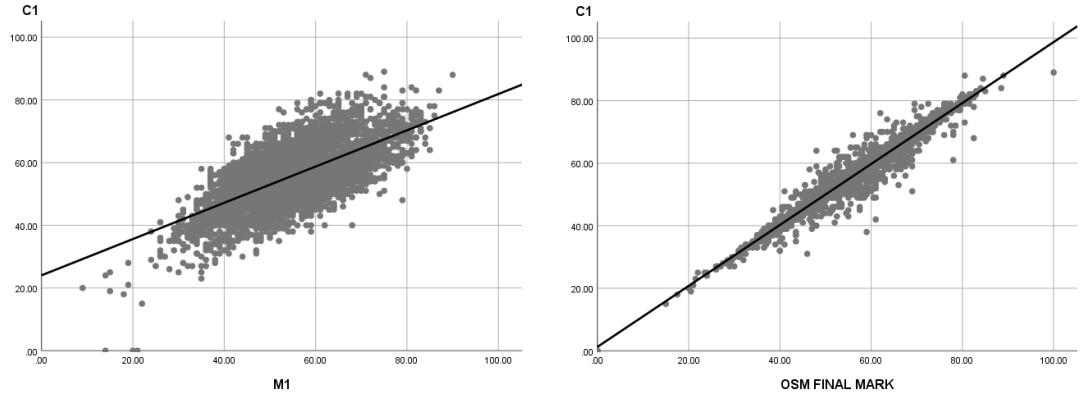
of data points appeared to be a bit smaller in Figure 2b.



(a) (b)

Figures 2a and 2b Regression lines of (a) M1 predicting C1 and (b) FM predicting C1 when only M1 contributed to the final marks

20. In contrast, when M1 did not make contribution to the final marks, R^2 dropped significantly to 0.438, an intercept increased to 24, and the slope coefficient dropped to 0.58. Whereas FM remained very close to C1, with a R^2 of 0.952, an intercept of only 1.26, and a slope of 0.975. As shown in regression lines, M1 data points scattered to form an oval shape in Figure 3a and the slope of the regression line was flat, while in Figure 3b, FM data points still tended to cluster along the regression line. This was the scenario where the need of double marking was most prominent.

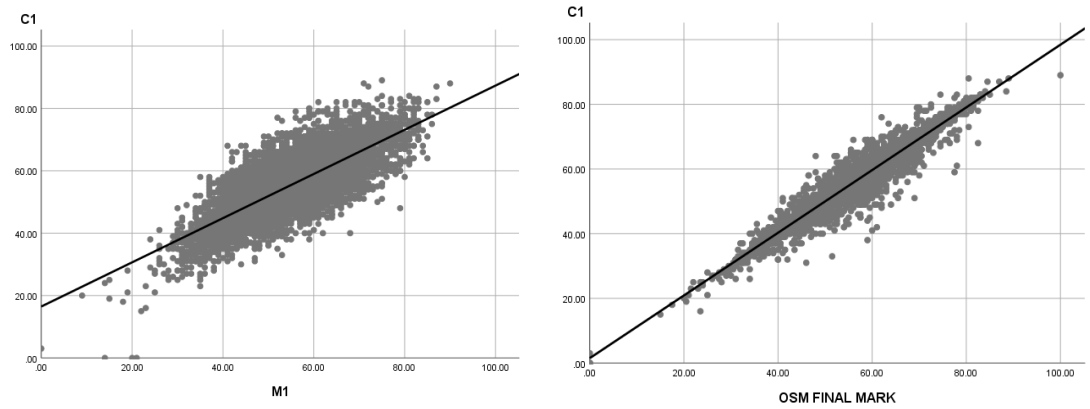


(a) (b)

Figures 3a and 3b Regression lines of (a) M1 predicting C1 and (b) FM predicting C1 when M1 did not contributed to the final marks

21. As expected, when all the 7195 cases were used to fit the regression model, the

results were in between the two scenarios discussed above. In general, FM was closer to "true scores" when compared with M1. As shown in Figures 4a and 4b, the spread of data points in Figure 4a was much larger than that in Figure 4b.



(a) (b)

Figures 4a and 4b Regression lines of (a) M1 predicting C1 and (b) FM predicting C1 of the whole data set

22. The conclusion drawn from the comparison of the above regression models was that FM was closer to the "true scores" than M1 in all scenarios. In other words, double marking was an effective mechanism in safeguarding marking quality.

Useful indicators of markers' marking accuracy

23. This study also examined two factors that might indicate markers' marking quality. As shown in Table 4, in our sample, the majority of the cases were marked by more experienced markers. These cases tended to have a higher mean score and a higher standard deviation in comparison with the cases marked by less experienced markers. About half of the samples in this study were marked by good performing markers and the remaining half were marked by average performing markers. The cases marked by average performing markers reported higher mean scores and slightly higher standard deviation when compared with those marked by good performing counterparts. The contrast between experience and performance raised the question of whether they indicate marking accuracy.

Table 4 Descriptive Statistics of different categories of markers

	N	Mean	Std. Deviation	Minimum	Maximum
All cases	7195 (100%)	5.48	4.43	0	31
More Experienced	4455 (62%)	5.57	4.47	0	31
Less Experienced	2740 (38%)	5.33	4.36	0	24
Good Performance	3535 (49%)	5.31	4.42	0	31
Average Performance	3660 (51%)	5.64	4.44	0	27

24. To answer the question, a regression analysis was conducted to all 7195 cases by coding the abovementioned two factors as dummy independent variables, EXP and PER, and the dependent variable was the absolute difference between C1 and M1 scores ($C1M1_{ABSDIFF}$).

$$C1M1_{ABSDIFF} = a + bEXP + cPER$$

25. As both independent variables were categorical variables, the model was essentially the same as an ANOVA model. As shown in Table 5, both EXP and PER were statistically significant at 0.05 level and were included in the model.

Table 5 ANOVA statistics of regression model

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	394.990	2	197.495	10.086	0.000
Residual	140827.719	7192	19.581		
Total	141222.710	7194			

a. Dependent Variable: $C1M1_{ABSDIFF}$

b. Predictors: (Constant), PER, EXP

26. The model as expressed in unstandardized terms was as follows:

$$C1M1_{ABSDIFF} = 5.47 + 0.35EXP - 0.42PER$$

27. The negative sign of the slope coefficient of PER implied that good performance helped produce smaller deviation from reference mark, or in other words better marking accuracy. And the positive sign of slope coefficient of EXP implied that more experienced markers tended to give marks farther away from the reference C1 mark. One plausible explanation of this finding is that more experience does not guarantee better performance. So, in future selection of markers, priority should be given to previous performance than years of experience.

Conclusion

28. In short, this study confirmed that double marking did help improve marking accuracy when using C1 as the reference.

29. It was also found that markers' previous performance was a more important indicator than their experience in securing reliable marks.

Recommendations for future studies

30. It was recommended that similar double marking studies could be extended to other subjects because the HKDSE Chinese Language writing examination has a maximum possible score of 103 and efficiency of double marking may change for items with a small maximum score or more strictly defined marking guidelines. In addition, if data on other potential indicators of markers marking accuracy can be collected, an enhanced marker selection scheme can be proposed.