Appendix 2
Statistical Report of the First Formal Run (SHSK0812)

1. Basic information

Table 1.1

Test	SHSK0812	
Date	14 Dec 2008	
No. of Participants	1,326	
Parallel Papers	**0812A(am)**	**0812B(pm)**
Total	721	605

The formal run was carried out in PolyU as part of the Graduating Students' Language Proficiency Assessment (GSLPA) and the participants were mainly final year undergraduate students of the University. There were two sessions of this run, where one was held in the morning and the other in the afternoon. Two parallel test papers were used, one for each session.

2. Overall

Figure 2.1 Mean Scores of the Three Papers

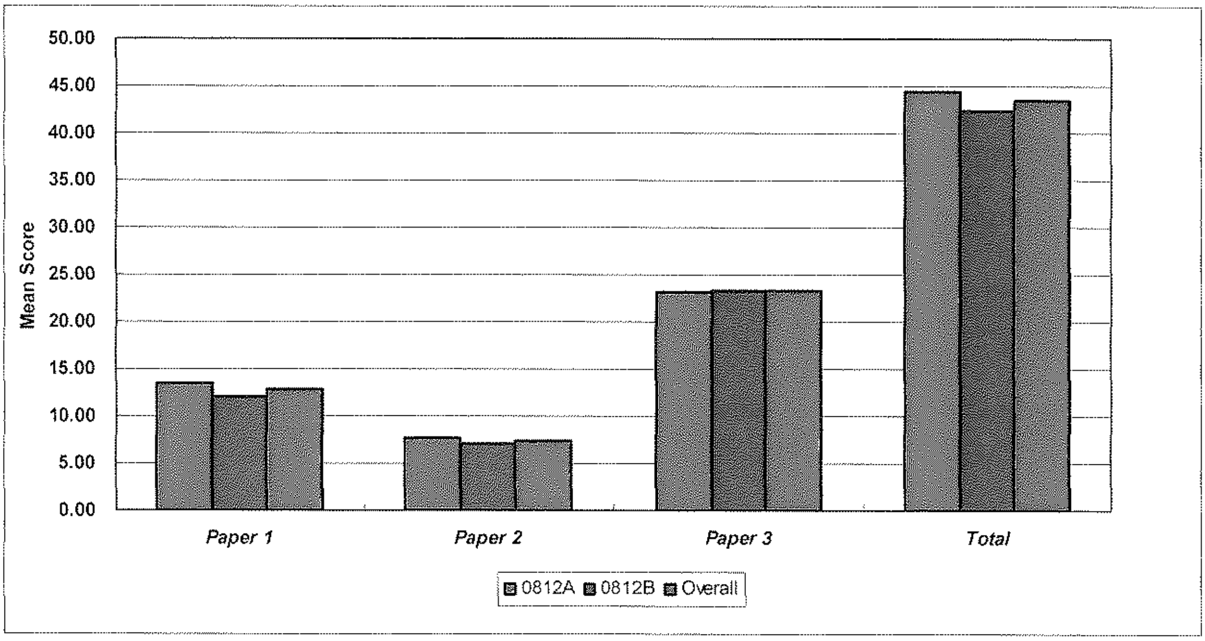


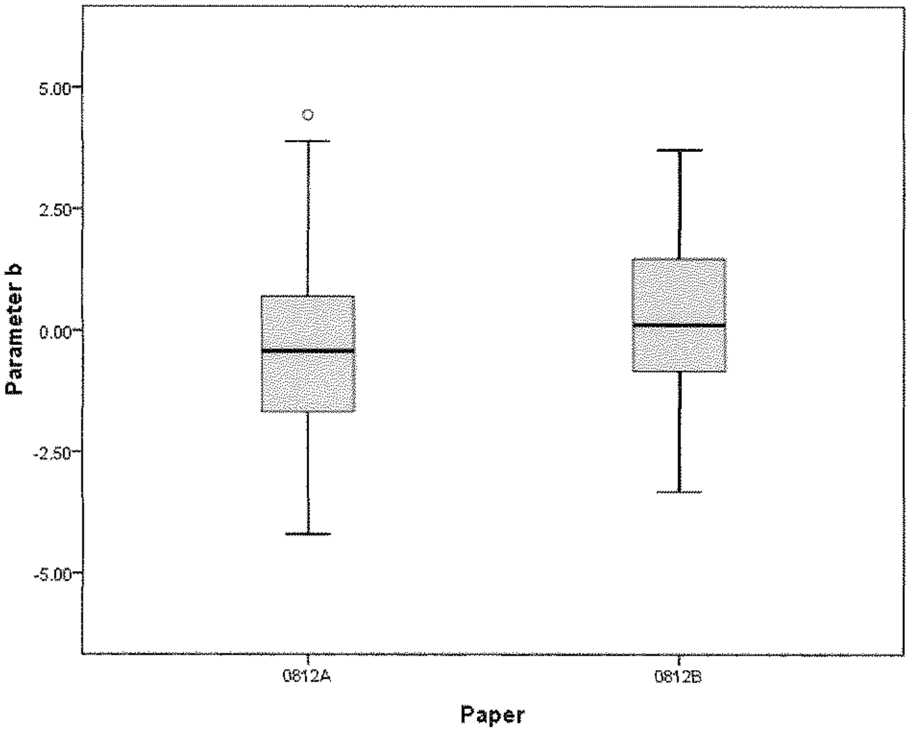
Figure 2.1 compares the mean scores of different parts of the test, and the results reveal that there is a high consistency between the two parallel papers used in the present run. It can be seen that the mean scores of paper 1-MC items (25%) are about 12-13, paper 2-cloze test (15%) around 7, paper 3-writing tasks (60%) about 23, and the overall (100%) between 42 to 45, which suggest that the test is of intermediate

level.

3. Paper 1: Multiple choice items (Basic knowledge and application principles of written Chinese)

In this paper, item response theory (IRT) is employed to analyze the characteristics of the items. Among different models of IRT, the three-parameter (3PL) model is adopted, where the estimation of item difficulty (parameter b), discrimination (parameter a) and the effect of guessing (parameter c) will be offered.

Figure 3.1 Item Difficulty (Parameter b)



From Figure 3.1, it can be seen that the median of item difficulty of both papers is about 0.0, which is near the center of the distribution, showing that this paper is a median-level test. Moreover, the box plots in Figure 3.1 indicate that the range of both papers is roughly between -4 and 4, revealing that the items are in a reasonable distribution by different degrees of difficulties. Figure 3.1 also demonstrates the consistency in terms of item difficulties across different papers, suggesting that the test is reliable.

Figure 3.2 Discrimination (Parameter a)

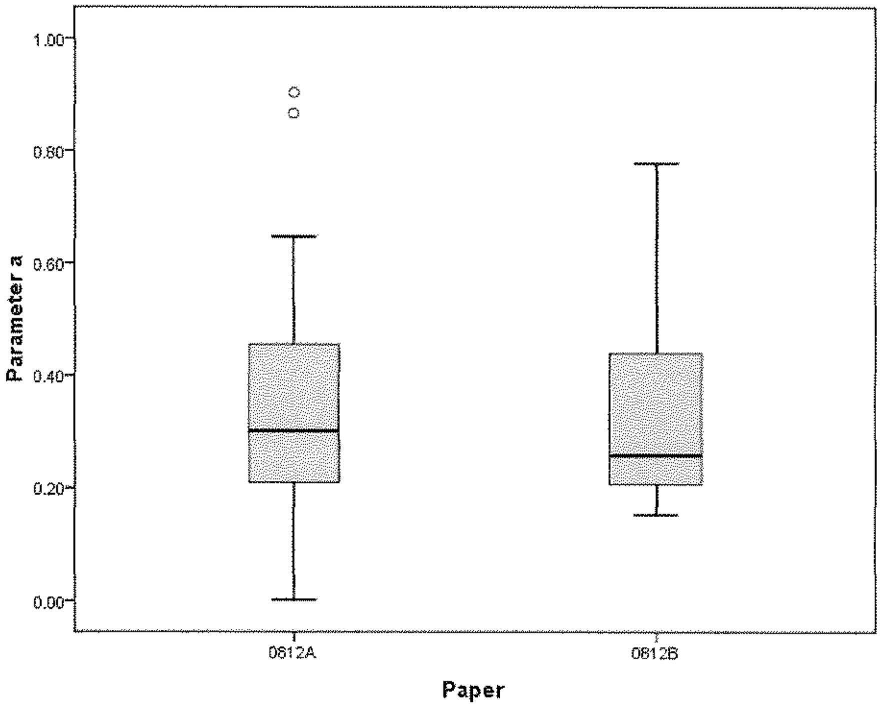


Figure 3.2 reports the discriminative power of the two papers. The medians of both of the two papers are about 0.3, attaining an acceptable level of discrimination. Comparatively, 0812B has a better distribution of items whose upper limit is approaching 0.8, a high level of discrimination. Despite the fact that the upper limit is lower than that of 0812B, 0812A has two items (displayed in Figure 3.2 by circles as outliers) exhibiting very good discriminative power (about 0.9).

Figure 3.3 The Effects of Guessing (Parameter c)

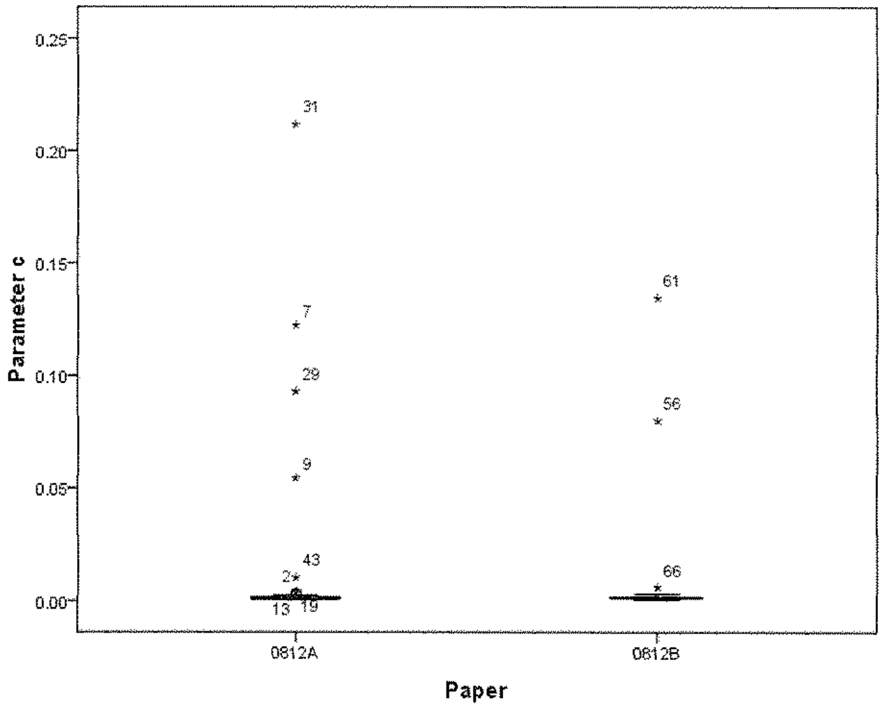


Figure 3.3 shows the effects of guessing on the probability of a correct response. It indicates the probability that very low ability individuals will get this item correct by chance. The extremely flat box plots in Figure 3.3 point to the fact that for almost all items, the probability of getting correct answers by chance is virtually zero. Even though there are a few extreme cases (expressed by asterisks), the most extreme one in 0812A is only about 0.22 and that in 0812B is about 0.14. In other words, the items of both papers are well-designed and the MC options function effectively so that it is not probable for a test-taker to get the correct answer by guessing.

4. Paper 2: The cloze test

Figure 4.1 Item Difficulty

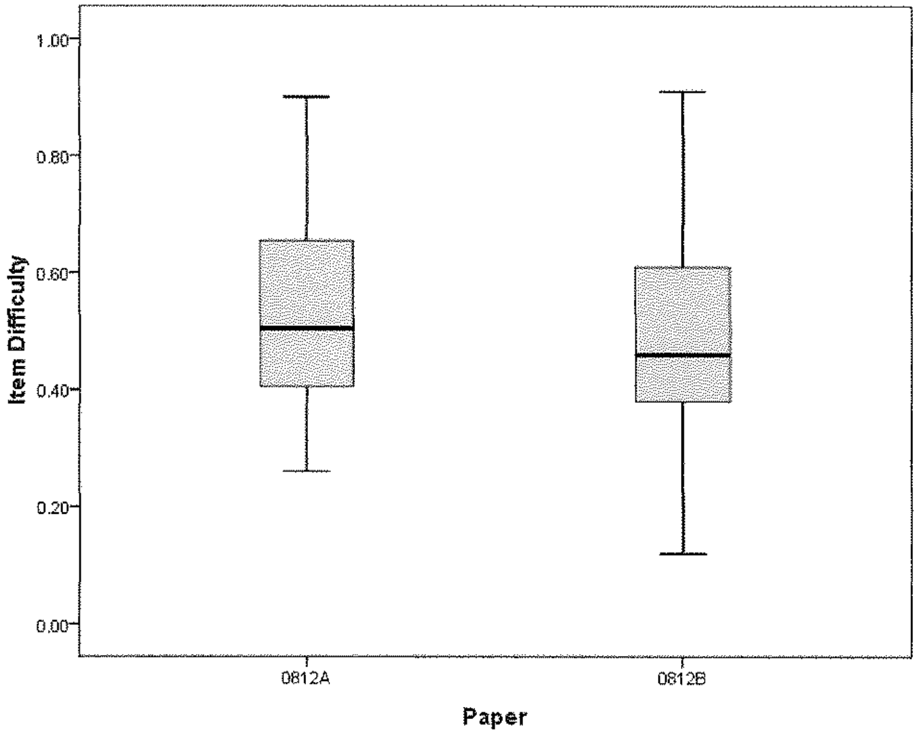


Figure 4.1 demonstrates a fine distribution of item difficulties – the medians of both papers are about 0.5, an intermediate level of difficulty, and the upper limits of both papers are about 0.9 while the lower limits are about 0.3 and 0.1 for 0812A and 0812B respectively, suggesting that the test papers consist items with different degree of difficulties, which are useful for assessing test-takers with different levels of proficiency.

Figure 4.2 Discrimination

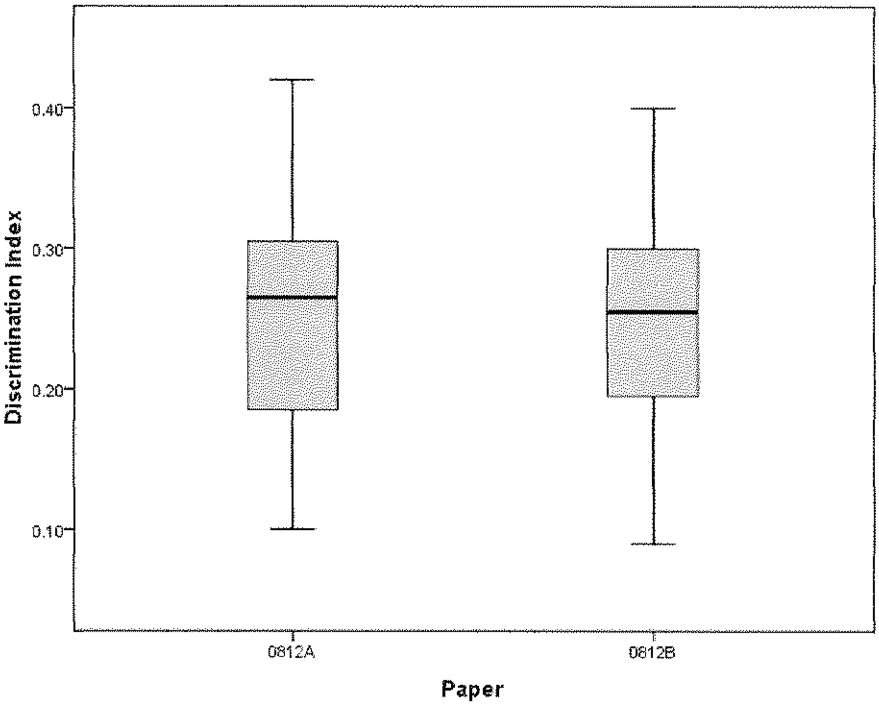
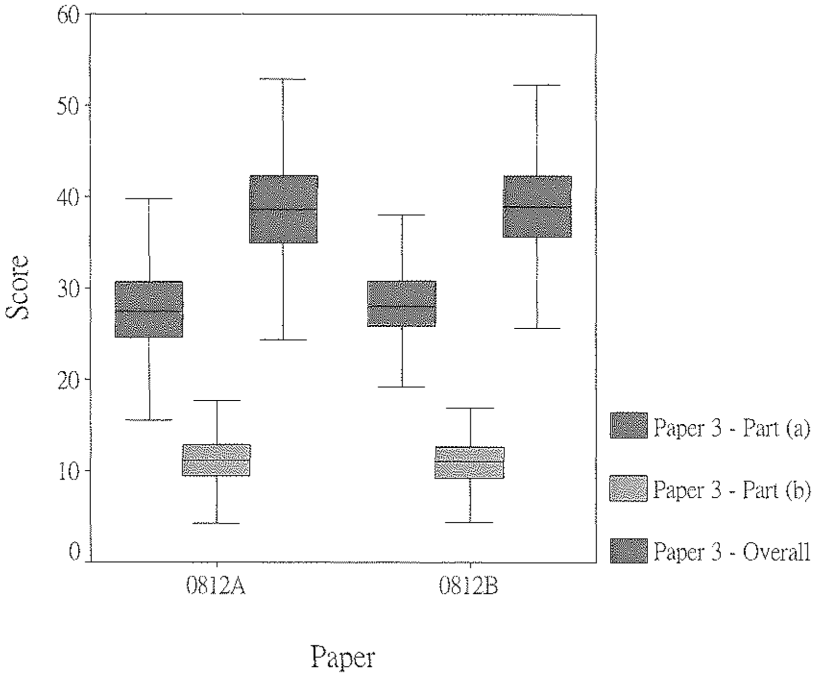


Figure 4.2 demonstrate the distribution of discrimination indexes. Overall speaking, the medians of both papers are above 0.25, attaining the threshold adopted in this paper. The Figure also reveals the distribution patterns of the two papers resemble to a large extent, showing that there is a consistency in terms of the discriminative power of the two papers.

5. Paper 3: Writing Tasks

Figure 5.1 Distributions of Scores



Appendix 2

Figure 5.1 compares the distributions of scores of part (a) – General writing, part (b) – Practical writing and overall in the two examination papers.

Firstly, in regard to the task of general writing (part (a)), as seen from the box plots in red, both papers show a similar amount of median scores. The range of scores of 0812A is a little wider than that of 0812B, yet the difference is not great.

Secondly, as far as the task of practical writing (part (b)) is concerned, papers 0812A and 0812B resemble each other, no matter in terms of medians or range of scores. Both of them have a median about 12 and range of scores from about 3 to 18.

Lastly, overall speaking, distributions of 0812A and 0812B present more or less the same pattern as well. From Figure 5.1, it is shown that the medians of papers are about 40 and they both range from about 25 to 55.

To conclude, the results demonstrate a high consistency between the writing tasks of the two parallel papers, which implies an effective control of the balance of task difficulty between the two papers as well as the steadiness in scoring, both are important considerations of subjective assessments like writing tasks.