Appendix 1

Statistical Report of the Pilot Tests[1]

1. Basic information

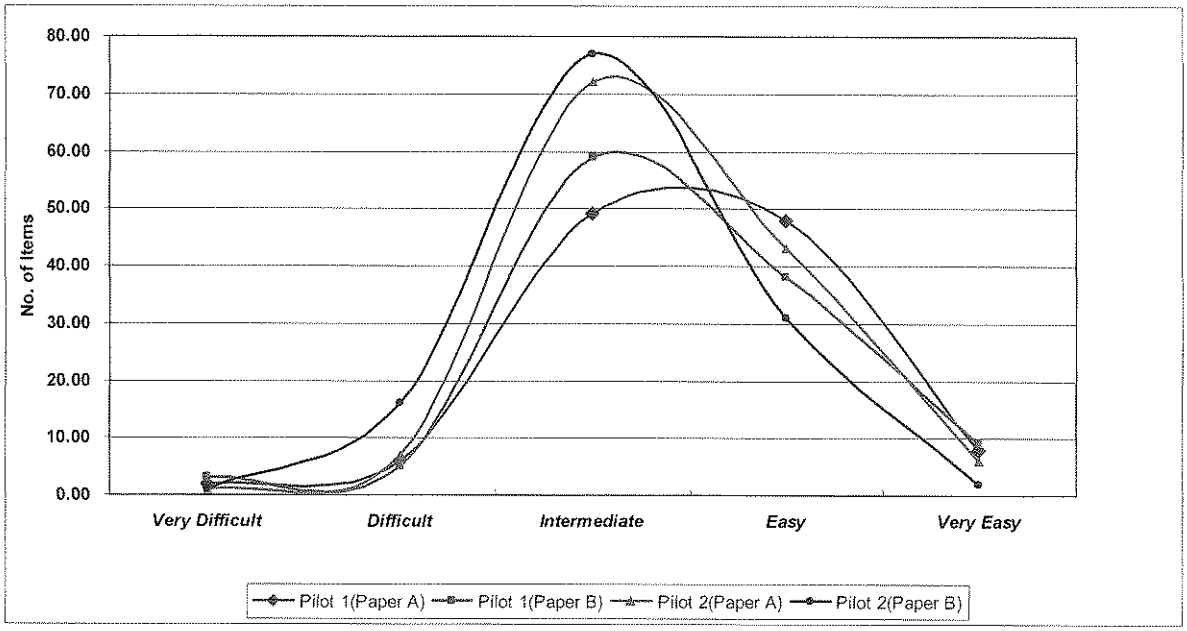
Table 1.1

Test	Pilot 1		Pilot 2	
Date	10-14 Nov 2007		30 Apr – 3 May 2008	
No. of Participants	100		126	
Parallel Papers	**A**	**B**	**A**	**B**
H	25	28	25	23
M	13	8	23	24
L	13	13	15	16
Total	**51**	**49**	**63**	**63**

The two pilots tests were carried out in PolyU and the participants were Year 1 undergraduate students of the University. They are sub-divided into high-level (H), medium-level (M) and low-level (L) groups according to the past SHSK results of the departments where the students belonged to. Two parallel test papers were used in each pilot test, and they were randomly assigned to the participants. As shown in Table 1.1, the number of participants in each paper of the same pilot test is more or less the same.

2. Overall

Figure 2.1 Distribution of Item Difficulty

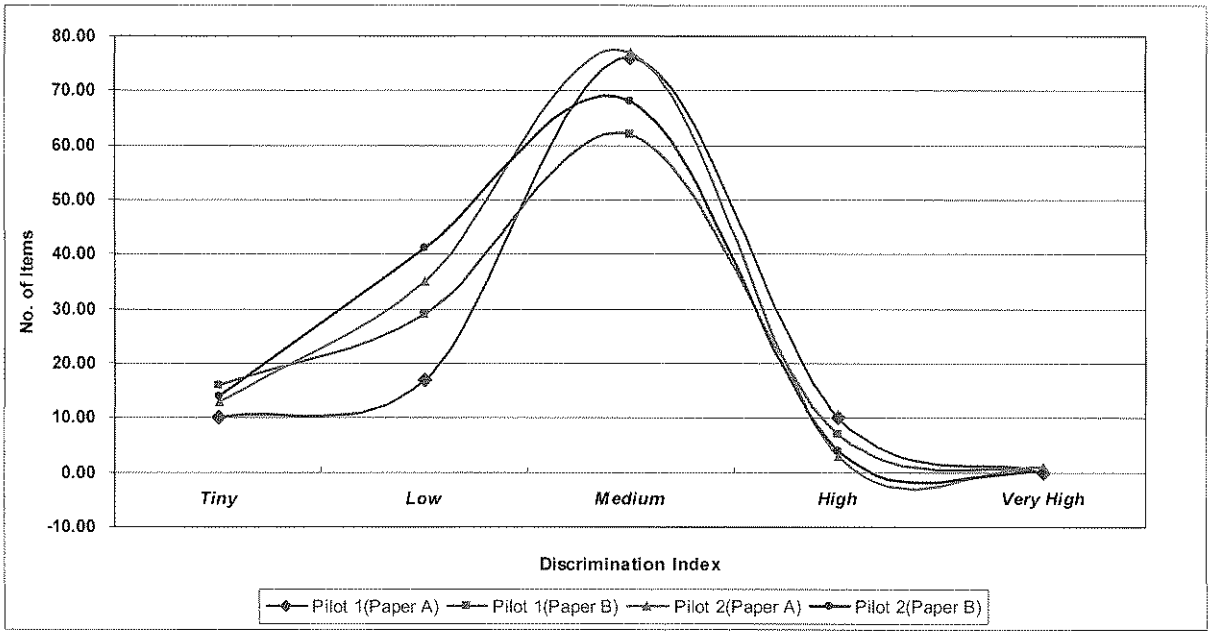


[1] In this report only objective items are included (i.e. MC items and the cloze test).

N.B: Very Difficult = 0.0-0.1; Difficult = 0.1-0.3; Intermediate = 0.3-0.7; Easy = 0.7-0.9; Very Easy = 0.9-1.0

Figure 2.1 indicates that the distribution of difficulties for all objective items in the SHSK test is more or less bell-shaped. Comparatively, the distributions of the two papers of pilot test 2 looks more like normal, in terms of both skewness and kurtosis, than that of pilot test 1, showing that the level of difficulty has been improved considerably.

Figure 2.2 Distribution of Item Discrimination



N.B: Tiny = 0.0-0.1; Low = 0.1-0.3; Medium = 0.3-0.7; High = 0.7-0.9; Very High = 0.9-1.0

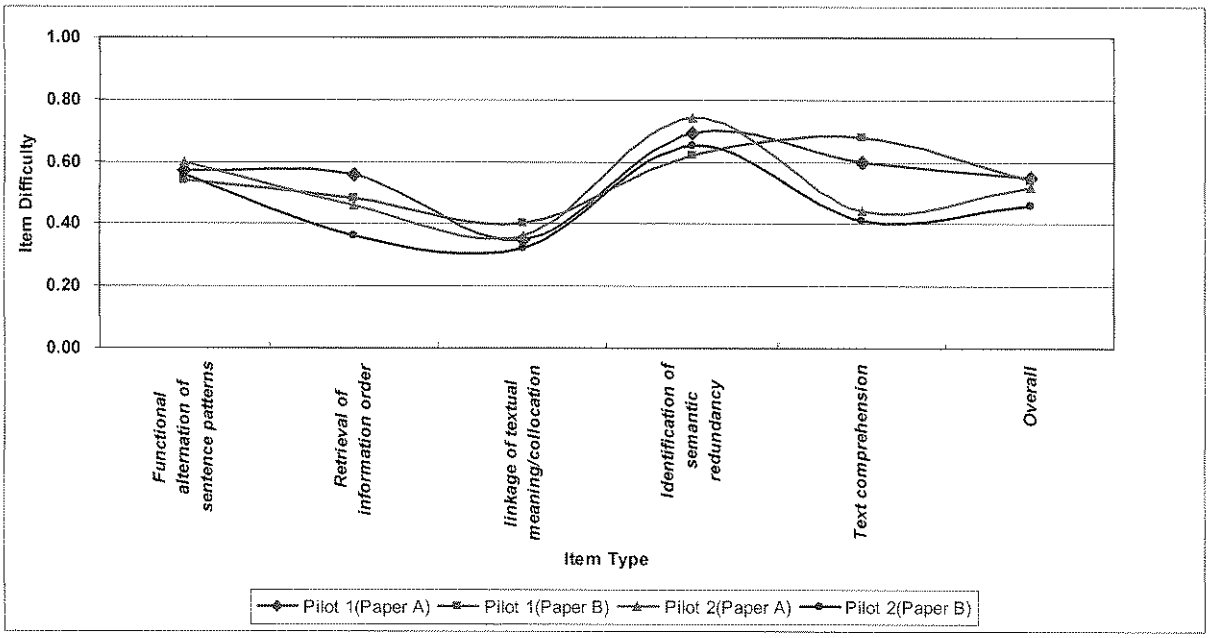
Figure 2.2 reveals that the distribution of discrimination indexes for all objective items in the SHSK test is more or less normal. In other words, most of the items in the papers are at least with medium discriminative power. It can also be seen that the distribution of different papers and different trials are to a large extent similar, suggesting a consistency in discriminative power of the test.

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

Table 3.1 Item Difficulty across Sections

Section	Pilot 1 (Paper A)	Pilot 1 (Paper B)	Pilot 2 (Paper A)	Pilot 2 (Paper B)
(1) Functional alternation of sentence patterns	0.57	0.54	0.60	0.56
(2) Retrieval of information order	0.56	0.48	0.46	0.36
(3) Linkage of textual meaning/collocation	0.35	0.40	0.36	0.32
(4) Identification of semantic redundancy	0.69	0.62	0.74	0.65
(5) Text comprehension	0.60	0.68	0.44	0.41
Overall	**0.55**	**0.54**	**0.52**	**0.46**

Figure 3.1 Item Difficulty across Sections



From Table 3.1 and Figure 3.1, it can be seen that the overall item difficulty is about 0.5, showing that this paper is a medium-level test, and item difficulties of different sections in this paper range from about 0.3 to 0.7, indicating 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 and trials, suggesting that the test is reliable.

Table 3.2 No. of Items with Discriminative Power[2]

Section	Pilot 1 (Paper A)	Pilot 1 (Paper B)	Pilot 2 (Paper A)	Pilot 2 (Paper B)
(1) Functional alternation of sentence patterns	8	7	10	8
(2) Retrieval of information order	8	7	9	7
(3) Linkage of textual meaning/collocation	10	9	9	9
(4) Identification of semantic redundancy	10	7	6	7
(5) Text comprehension	8	9	7	8
Overall	**44**	**39**	**41**	**39**

Figure 3.2 No. of Items with Discriminative Power

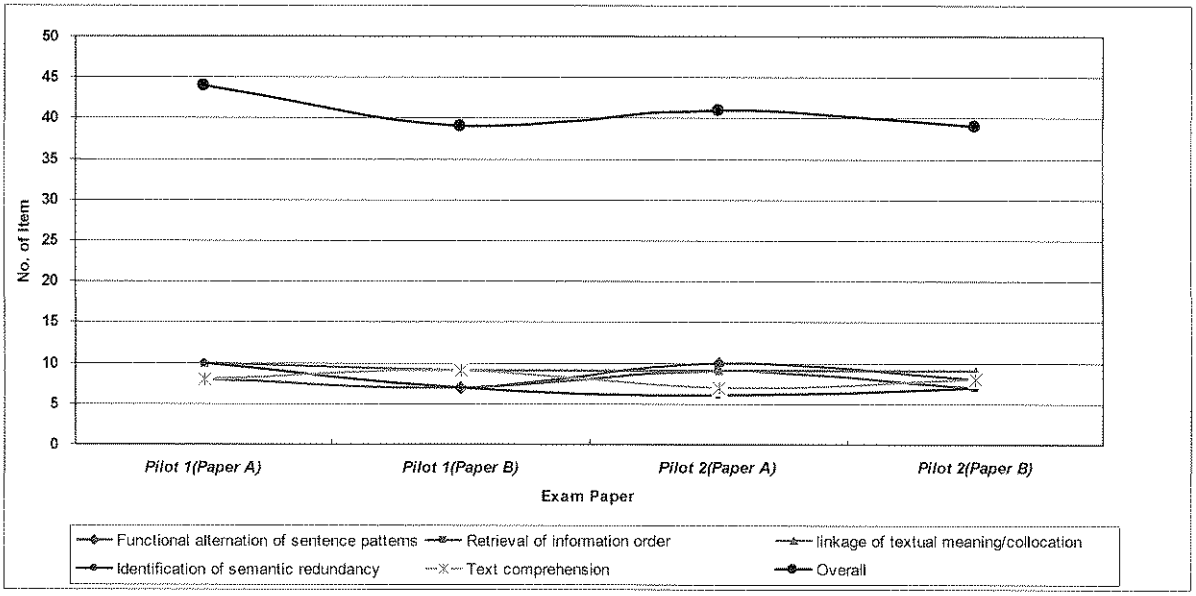


Table 3.2 and Figure 3.2 report the number of items that attained an acceptable level of discrimination. Overall speaking, there are about 80-90% of items being acceptable, and the acceptable rates in various sections of the paper are more or less the same, showing that the test is of good discriminative power.

[2] The item discrimination is measured by point-biserial correlation and the threshold of acceptance is 0.3.

4. Paper 2: The cloze test

Table 4.1 Item Difficulty across Articles

Article	Pilot 1 (Paper A)	Pilot 1 (Paper B)	Pilot 2 (Paper A)	Pilot 2 (Paper B)
Argumentative	0.69	0.73	0.69	0.67
Expository	0.70	0.62	0.71	0.62
Literary	0.72	0.66	0.63	0.63
Overall	**0.70**	**0.67**	**0.68**	**0.64**

Figure 4.1 Item Difficulty across Articles

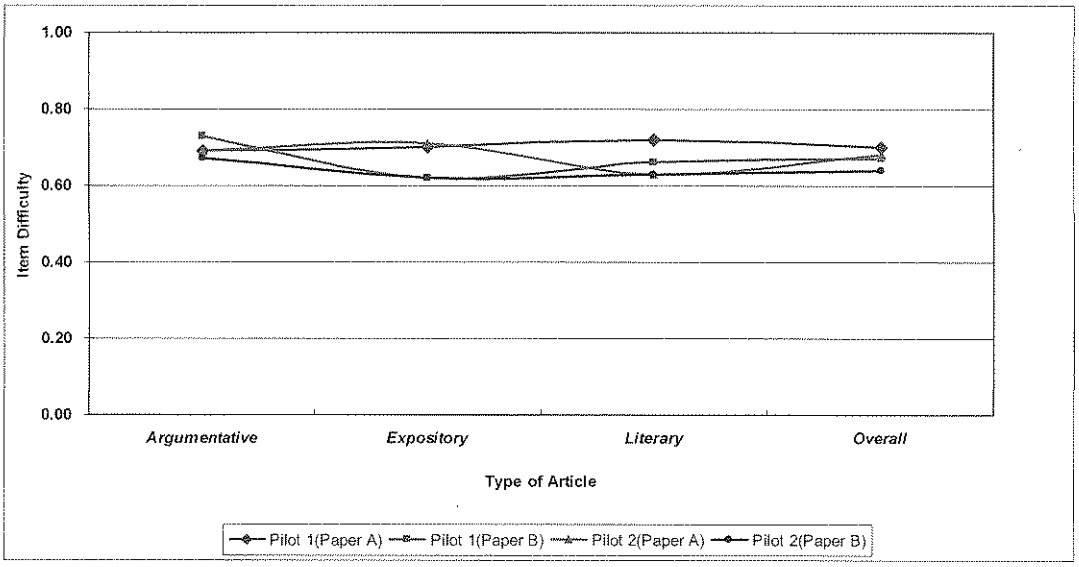


Table 4.1 and Figure 4.1 demonstrate a highly consistent distribution of item difficulties, no matter in terms of articles or different papers and trials. This reveals that the test is reliable in regard to its degree of difficulty. From Table 4.1 and Figure 4.1, it can also be seen that on the whole, the level of item difficulty is around 0.7, showing that the test is manageable by most of the test-takers.

Table 4.2 No. of Items with Discriminative Power[3]

Article	Pilot 1 (Paper A)	Pilot 1 (Paper B)	Pilot 2 (Paper A)	Pilot 2 (Paper B)
Argumentative	12	10	10	8
Expository	12	14	19	14
Literary	17	7	11	11
Overall	**43**	**31**	**40**	**33**

[3] The item discrimination is measured by Pearson product-moment correlation and the threshold of acceptance is 0.25.

Figure 4.2 No. of Items with Discriminative Power

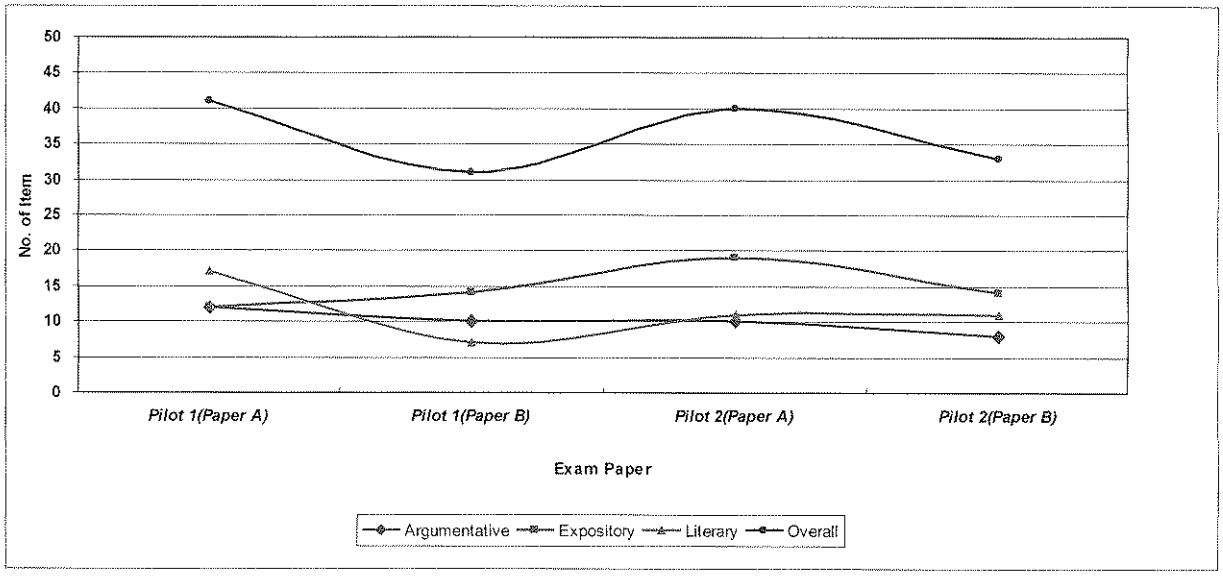


Table 4.2 and Figure 4.2 demonstrate the number of items that attained an acceptable level of discrimination. Overall speaking, there are 30-40 of items being acceptable, and the number of acceptable items in different types of article varies. Among them, it can be seen that the literary ones are rather unstable in terms of discriminative power; therefore this genre has been excluded in the formal run.