Research Note: Effects of Skill Integration on Language Assessment: A Comparative Study of Pearson Test of English Academic and Internet-Based College English Test Band-6

Yan JIN
Shanghai Jiao Tong University, China

Xiaoyi ZHANG
Shanghai Jiao Tong University, China

February 2014

1. Introduction

The study aims to investigate whether the improvement in task authenticity using integrated tasks would be accompanied by muddied construct validity in language assessment. The issue was explored through a comparative study between Pearson Test of English Academic (PTE Academic) and Internet-Based College English Test-Band 6 (IB-CET6). In the study, "integration" is defined as an integrative notion which includes the synthetic operation of both language skills and information (Zahedi & Shamsaee, 2012). The interpretation of "skill integration", accordingly, pertains not only to language skills but also to skills in manipulating information in an integrative manner.

PTE Academic is an innovative language assessment using mainly integrated tasks, whereas IB-CET6 uses mainly independent tasks of listening, speaking, reading, and writing. Both tests, however, follow the tradition of profiling test takers' performances on the four communicative skills. In the study, the influence of task type on a test's construct validity will be explored by seeking answers to the following three research questions: 1) How comparable are the two tests in rank-ordering test takers? 2) How do test takers perceive the difficulty and effectiveness of the tasks and tests? 3) How and to what extent do test takers' cognitive processes and use of metacognitive strategies differ when engaged in integrated and independent tasks?

2. Methodology

2.1 Data collection

Following a pilot study (n=16) conducted in a university in Nanjing, China, 43 volunteers in universities in Shanghai and Beijing took the live tests of PTE Academic and IB-CET6, within one month between May and July 2012. After each test, the test takers completed questionnaires on test and task evaluation and processes of task completion. The questions were drafted based on context validity and theory-based validity (cognitive validity) of the socio-cognitive framework for validating language tests (Weir, 2005). Other sources we referred to when drafting the questionnaires include the physical and linguistic features of spoken language (Rost, 2002:31, 171-2, cited in Weir, 2005:99), studies of integrated writing (e.g., Friend, 2001; Garner, 1982; Kennedy, 1985; Yang & Shi, 2003), and reading strategies such as skimming for gist, scanning, and search

reading (Urquhart & Weir, 1998). The questionnaires for the two tests have parallel structures but slightly different questions to elicit respondents' views on 1) the difficulty of the two tests at the section and the test levels, 2) performance-related factors such as familiarity with the test content and format, anxiety and fatigue during the test, and interface used for taking the test, 3) cognitive processing in the two tests, and 4) major differences between the two tests and ways to improve the tests. The five-point Likert scale was used for all survey questions, except the open ones.

2.2 Data analysis

Data were analyzed using SPSS 19.0. The data of 10 candidates were considered invalid, mainly due to their missing or abnormal patterns of responses in the survey (e.g., missing responses larger than 5% or selection of 1 or 5 throughout one part of the survey questionnaire). To compare the performances of the candidates at different proficiency levels, the 33 participants were divided into high-scorer (the top one-third), mid-scorer (the middle one-third), and low-scorer (the bottom one-third) groups. The number of test takers in each group varied from 10 to 12. ANOVA and post-hoc analyses of the scores confirmed significant differences among the groups and between each pair of the groups (p=.00, F=31.06~79.74). One-way ANOVA analyses, together with post-hoc comparisons using the Bonferroni test, were conducted to compare how high-, mid- and low-scorer groups performed in independent and integrated tasks. Survey data were mainly analyzed using paired sample t-test. Given that the sample size was small, the effect size of each elicited item was also calculated for further examination.

3. Results

3.1 Test score analysis: rank-ordering comparability

Though the sample size was small, test scores of the participants achieved reasonably satisfactory distributions. The skewness value of the total and subscores ranges from -1.22 (the speaking section in IB-CET6) to 1.36 (the reading section in PTE Academic) with no significant deviation from zero, allowing for further statistical analyses.

3.1.1 Correlational analyses

Correlational analyses of the total and subscores of the two tests were conducted to investigate their comparability in rank-ordering test takers (Table 1). Total scores of the tests correlate quite satisfactorily (r=.84), providing evidence for the comparability of the two tests in rank-ordering test takers in terms of their overall proficiency. Scores of the four skills correlate moderately (r=.51~.72), which raises some concern over the comparability at the level of component skills. PTE Academic has much higher internal correlations (r=.58~.90) than IB-CET6 (r=.22~.61).

Table 1: Correlations of PTE Academic and IB-CET6 test scores

	PTE Academic					IB-CET6				
	T	L	S	R	W	T	L	S	R	W
PTE Academic_T	1	.92**	.87**	.94**	.88**	.84**	.73**	.50**	.64**	.71**
PTE Academic_L		1	.70**	.79**	.90**	.86**	.72**	.44*	.74**	.66**
PTE Academic_S			1	.83**	.58**	.67**	.55**	.51**	.48**	.60**
PTE Academic_R				1	.77**	.80**	.70**	.48**	.59**	.71**
PTE Academic_W					1	.72**	.66**	.35*	.55**	.63**
IB-CET6_T						1	.89**	.58**	.80**	.74**
IB-CET6_L							1	.53**	.55**	.61**
IB-CET6_S								1	.22	.32
IB-CET6_R									1	.43*
IB-CET6_W										1

Notes. T=total, L=listening, S=speaking, R=reading, W=writing; *p<.05, **p<.01.

3.1.2 Cross-tab analyses

To further investigate whether the test takers were categorized into the same proficiency groups (i.e., high-, mid- and low-scorer groups) by their performances on the two tests, cross-tabulations of the total and subscores were performed (Table 2). Test takers in those highlighted boxes were mismatched because they were categorized as a high-scorer in one test but a low-scorer in the other. A further check of the identity of the 14 mismatched test takers in bold showed that 11 test takers, one-third of the whole group, were mismatched for their overall performance and/or their performance on one or more components in the two tests.

Table 2: Cross-tabulation of PTE Academic and IB-CET6 high-, mid-, low-scorers

		IB-CET6														
		Total			Listening			Speaking			Reading			Writing		
		L	M	H	L	M	H	L	M	H	L	M	H	L	M	H
PTE Academic	L	7	2	**2**	5	4	**2**	7	4	**0**	6	2	**3**	6	6	**0**
	M	4	5	2	5	4	2	2	4	5	4	6	1	2	4	4
	H	**0**	4	7	**1**	3	7	**2**	4	5	**2**	3	6	**2**	2	7

Notes. L=low-scorer group, M=mid-scorer group, H=high-scorer group.

3.2 Survey data analysis: test takers' evaluation of the tests and tasks

3.2.1 Difficulty level of the tests and tasks

Means of the questions on the difficulty level of the tests and component sections were calculated, and paired sample t-tests were conducted, to compare the differences (Table 3). Both tests were considered quite difficult and the means of the overall difficulty of the tests were exactly the same (2.39). There were no significant differences between the means of the listening, speaking and reading sections of the tests, all being considered quite challenging with a mean lower than 3.00. The writing section of IB-CET6, the only section with a mean over 3.00, was perceived as significantly easier than the writing section of PTE Academic (p=.00).

Table 3: Mean differences of questions on difficulty level of tests and sections

	IB-CET6		PTE Academic		MD	Sig.
	Mean	SD	Mean	SD		
Listening	2.47	.95	2.28	.58	.19	.23
Speaking	2.88	.89	2.67	.69	.21	.32
Reading	2.88	1.04	2.50	.80	.38	.09
Writing	3.36	.82	2.58	.83	.78	.00
Overall test	2.39	.96	2.39	.88	.00	1.00

Notes. 1=very difficult, 5=very easy; SD=Standard deviation; MD=Mean difference (IB-CET6–PTE Academic); p<.05 (two-tailed).

Individual tasks of the two tests were also considered difficult (task-level data were not included in this summary report). IB-CET6 has three tasks with a mean over 3.00: *multiple-choice skimming and scanning* (M=3.06), *writing a summary* (M=3.06) and *writing an essay* (M=3.21). PTE Academic also has three tasks with a mean equalling or higher than 3.00: *read aloud* (M=3.42), *selecting missing words* (M=3.00) and *highlighting incorrect words* (M=3.36) in the listening section. The listening task of *passage comprehension* in IB-CET6 (M=2.48) and the listening task of *summarizing spoken text* in PTE Academic (M=2.15) were perceived as the most difficult task of each test.

3.2.2 Effectiveness of the tests and tasks

The effectiveness of IB-CET6 in assessing English abilities was perceived more favorably than PTE Academic, both at the test and the task levels, suggesting test takers' preference for the traditional, independent tasks adopted in IB-CET6. Paired sample t-tests identified ten questions which have a significant mean difference, with a medium to large effect size (Table 4). For all these questions, IB-CET6 got a higher mean than PTE Academic.

Table 4: Mean differences of questions on test evaluation

Questions on test evaluation	IB-CET6	PTE Academic	MD	Sig.	*d*
1. Clarity of task description	3.61	2.94	.67	.01	.62
4. Task type: reading	3.22	2.63	.59	.01	.61
5. Task type: writing	3.45	2.97	.48	.03	.57
9. Time allowed for listening tasks	3.61	3.00	.61	.02	.60
10. Time allowed for speaking tasks	3.36	2.58	.78	.01	.70
11. Time allowed for reading tasks	3.42	2.73	.69	.01	.57
12. Time allowed for writing tasks	3.82	2.97	.85	.00	.94
16. Effectiveness of writing tasks	3.75	3.25	.50	.01	.64
22. Arrangement of test time	3.50	2.81	.69	.00	.72
28. Seldom felt anxious and nervous	3.39	2.85	.54	.03	.44

Notes. The larger the value, the more positive the evaluation. 1=absolutely disagree, 5=absolutely agree; MD=Mean difference (IB-CET6–PTE Academic); p<.05 (two-tailed); d=Cohen's (1988) *d,* d=.2 (small), d=.5 (medium), d=.8 (large).

3.3 Survey data analysis: cognitive processes and strategies

3.3.1 A comparison of high-, mid- and low-scorer groups (ANOVA)

ANOVA analyses identified 13 questions with significant differences among the high-, mid- and low-scorer groups in the listening and reading sections of the two tests and the writing section of IB-CET6 (η^2=.19~.34), but not in the speaking sections of the two tests, or the writing section of PTE Academic. The results are specified as below.

Listening (Tables 5, 6, 7)

In the listening section of IB-CET6, the strategy of *I read the questions in order to know what I should focus on while listening* (p=.02, η^2=.22) and the factor of *speaker's gender* (p=.04, η^2=.21) had statistically significant differences. Particularly, post-hoc tests showed that the high-scorer group (M=4.27) was more skilled in setting goals for listening tasks than the low-scorer group (M=3.09). In PTE Academic, the only item with a significant difference is *taking notes to facilitate memorizing and understanding* (p=.03, η^2=.22), which was more frequently adopted by high-scorers (M=4.00) than low-scorers (M=3.00).

Table 5: Questions with significant differences between groups: listening tasks

		Sum of Squares	df	Mean Square	F	Sig.	η^2
1. IB-CET6: I read the questions in order to know what I should focus on while listening.	Between Groups	8.42	2	4.21	4.34	.02	.22
	Within Groups	29.09	30	.97			
	Total	37.52	32				
2. IB-CET6: Factors affecting my performance: Gender of the speaker (male/female).	Between Groups	7.50	2	3.75	3.79	.04	.21
	Within Groups	28.72	29	.99			
	Total	36.22	31				
3. PTE Academic: I took notes to facilitate my memorizing and understanding.	Between Groups	5.64	2	2.82	4.12	.03	.22
	Within Groups	20.55	30	.69			
	Total	26.18	32				

Notes. p<.05; η^2=.01 (small), η^2=.06 (medium), η^2=.14 (large) (Larson-Hall, 2010).

Table 6: Descriptives of the groups: listening tasks

		N	Mean	SD
1. IB-CET6: I read the questions in order to know what I should focus on while listening.	low	11	3.09	1.04
	mid	11	4.00	1.00
	high	11	4.27	.91
2. IB-CET6: Factors affecting my performance: Gender of the speaker (male/female).	low	11	2.82	.87
	mid	11	1.73	1.10
	high	10	1.90	.99
3. PTE Academic: I took notes to facilitate my memorizing and understanding.	low	11	3.00	.63
	mid	11	3.64	.92
	high	11	4.00	.89

Note. SD=Standard deviation.

Table 7: Multiple comparisons between groups (Bonferroni): listening tasks

Dependent Variable	(I) Scorer Groups	(J) Scorer Groups	Mean Difference (I-J)	Std. Error	Sig.
1. IB-CET6: I read the questions in order to know what I should focus on while listening.	low	Mid	-0.91	.42	.12
		High	-1.18*	.42	.03
	mid	Low	.91	.42	.12
		High	-0.27	.42	1.00
	high	Low	1.18*	.42	.03
		Mid	.27	.42	1.00
2. IB-CET6: Factors affecting my performance: Gender of the speaker (male/female).	low	Mid	1.09	.42	.05
		High	.92	.44	.13
	mid	Low	-1.09	.42	.05
		High	-0.17	.44	1.00
	high	Low	-0.92	.44	.13
		Mid	.17	.44	1.00
3. PTE Academic: I took notes to facilitate my memorizing and understanding.	low	Mid	-0.64	.35	.24
		High	-1.00*	.35	.02
	mid	Low	.64	.35	.24
		High	-0.36	.35	.93
	high	Low	1.00*	.35	.02
		Mid	.36	.35	.93

Note. *p<.05.

Reading (Tables 8, 9, 10)

The facet of *time management* in in-depth reading of IB-CET6 (p=.00, η^2=.33) and *the logical relationship between sentences and paragraphs* in reading passages of PTE Academic (p=.01, η^2=.30) were found to be statistically significant. In IB-CET6, the high-scorer group (M=4.56) allocated time more effectively than both the mid-scorers (M=3.55) and low-scorers (M=3.17). In PTE Academic, the mean differences between the high- (M=3.45) and low-scorer groups (M=2.82), as well as the mid- (M=3.45) and low-scorer groups are also significant, indicating a proportional relationship between language proficiency and the logical awareness at the clausal and textual levels.

Table 8: Questions with significant differences between groups: reading tasks

		Sum of Squares	df	Mean Square	F	Sig.	η^2
1. IB-CET6: In-depth reading: I had a good control of time for all questions.	Between Groups	10.26	2	5.13	7.22	.00	.33
	Within Groups	20.62	29	.71			
	Total	30.88	31				
2. PTE Academic: Factors affecting my performance: Logical relationship between sentences/paragraphs.	Between Groups	2.97	2	1.49	6.28	.01	.30
	Within Groups	7.09	30	.24			
	Total	10.06	32				

Notes. p<.05; η^2=.01 (small), η^2=.06 (medium), η^2=.14 (large) (Larson-Hall, 2010).

Table 9: Descriptives of the groups: reading tasks

		N	Mean	SD
1. IB-CET6: In-depth reading: I had a good control of time for all questions.	low	12	3.17	1.12
	mid	11	3.55	.69
	high	9	4.56	.53
2. PTE Academic: Factors affecting my performance: Logical relationship between sentences/paragraphs.	low	11	2.82	.41
	mid	11	3.45	.52
	high	11	3.45	.52

Note. SD=Standard deviation.

Table 10: Multiple comparisons between groups (Bonferroni): reading tasks

Dependent Variable	(I) Scorer Groups	(J) Scorer Groups	Mean Difference (I-J)	Std. Error	Sig.
1. IB-CET6: In-depth reading: I had a good control of time for all questions.	low	Mid	-0.38	.35	.88
		High	-1.39*	.37	.00
	mid	Low	.38	.35	.88
		High	-1.01*	.38	.04
	high	Low	1.39*	.37	.00
		Mid	1.01*	.38	.04
2. PTE Academic: Factors affecting my performance: Logical relationship between sentences/paragraphs.	low	Mid	-0.64*	.21	.01
		High	-0.64*	.21	.01
	mid	Low	.64*	.21	.01
		High	.00	.21	1.00
	high	Low	.64*	.21	.01
		Mid	.00	.21	1.00

Note. *p<.05.

Writing (Tables 11, 12, 13)

In integrated writing (writing a summary), IB-CET6 had eight questions with a statistical significance related to the preparatory phase (*Q1: summary-while reading the passage, I paid attention to the key information it contains*; *Q2: summary-while reading the passage, I associated the passage with my previous knowledge*) and revising phase (*Q3: summary-after writing, I examined if my personal opinions had been included into the summary*; *Q4: summary-after writing, I examined if my summary was coherent*; *Q5: summary-after writing, I refined my language*). Post-hoc comparisons showed that for *Q1*, the mean score of the high-scorer group (M=4.64) was significantly higher than both low- (M=3.50) and mid-scorer groups (M=3.50), suggesting the attention paid by the high-scorers to identifying key information in the reading passage in summary-writing tasks. No significant difference was observed between any two of the scorer groups for *Q2*. Significant differences in the revising phase were identified between the low- and high-scorer groups. In IB-CET6, the high-scorers seemed to have performed a more complete writing process than the low scorers: the high-scorers paid more attention to revising activities including checking relevance and comprehensiveness of the content and cohesion of the text, and polishing the language.

In independent writing (writing an essay), no questions with statistically significant differences were identified in PTE Academic. For IB-CET6, time management proved to be an important aspect of the essay-writing task. The mid-scorer group (M=4.00) reported that they paid more attention at the planning stage to *allocating time for each stage of writing* (planning, writing and revision) than the high-scorer group (M=2.82). The high-scorers (M=4.36), interestingly, agreed more strongly than the low- (M=3.22) and mid-scorer

groups (M=3.33) to the statement that *time allowed for the essay writing task* is a factor affecting writing performance, as indicated by the post-hoc comparison of *Q7*. Finally, the mid-scorer group (M=2.75) found it more difficult to *organize their writing properly* than did the high-scorer group (M=3.55).

Table 11: Questions with significant differences between groups: writing tasks

		Sum of Squares	df	Mean Square	F	Sig.	η^2
1. IB-CET6 Summary: While reading the passage, I paid attention to the key information it contains.	Between Groups	9.47	2	4.74	7.87	.00	.34
	Within Groups	18.05	30	.60			
	Total	27.52	32				
2. IB-CET6 Summary: While reading the passage, I associated the passage with my previous knowledge.	Between Groups	8.23	2	4.12	3.47	.04	.19
	Within Groups	35.65	30	1.19			
	Total	43.88	32				
3. IB-CET6 Summary: After writing, I examined if my personal opinions had been included into the summary.	Between Groups	10.09	2	5.04	3.78	.03	.20
	Within Groups	39.98	30	1.33			
	Total	50.06	32				
4. IB-CET6 Summary: After writing, I examined if my summary was coherent.	Between Groups	7.10	2	3.55	3.83	.03	.20
	Within Groups	27.81	30	.93			
	Total	34.91	32				
5. IB-CET6 Summary: After writing, I refined my language.	Between Groups	11.89	2	5.95	5.95	.01	.28
	Within Groups	29.99	30	1.00			
	Total	41.88	32				
6. IB-CET6 Essay: Before writing, I planned the procedure and allocated time for each stage.	Between Groups	8.31	2	4.15	4.13	.03	.22
	Within Groups	29.19	29	1.01			
	Total	37.50	31				
7. IB-CET6 Essay: Factors affecting my performance: time allowed for writing.	Between Groups	8.45	2	4.23	5.90	.01	.29
	Within Groups	20.77	29	.72			
	Total	29.22	31				
8. IB-CET6 Essay: Level of difficulty to meet the requirements: organizing the essay properly.	Between Groups	3.67	2	1.83	3.77	.04	.20
	Within Groups	14.58	30	.49			
	Total	18.24	32				

Notes. $p<.05$; $\eta^2=.01$ (small), $\eta^2=.06$ (medium), $\eta^2=.14$ (large) (Larson-Hall, 2010).

Table 12: Descriptives of the groups: writing tasks

		N	Mean	SD
1. IB-CET6 Summary: While reading the passage, I paid attention to the key information it contains.	low	10	3.50	.97
	mid	12	3.50	.80
	high	11	4.64	.51
2. IB-CET6 Summary: While reading the passage, I associated the passage with my previous knowledge.	low	10	2.70	.68
	mid	12	2.50	1.00
	high	11	3.64	1.43
3. IB-CET6 Summary: After writing, I examined if my personal opinions had been included into the summary.	low	10	2.60	1.17
	mid	12	3.67	.89
	high	11	3.91	1.38
4. IB-CET6 Summary: After writing, I examined if my summary was coherent.	low	10	3.20	1.23
	mid	12	3.83	.58
	high	11	4.36	1.03
5. IB-CET6 Summary: After writing, I refined my language.	low	10	2.70	1.16
	mid	12	3.25	.75
	high	11	4.18	1.08
6. IB-CET6 Essay: Before writing, I planned the procedure and allocated time for each stage.	low	9	3.22	.83
	mid	12	4.00	.95
	high	11	2.82	1.17
7. IB-CET6 Essay: Factors affecting my performance: time allowed for writing.	low	9	3.22	.67
	mid	12	3.33	.99
	high	11	4.36	.81
8. IB-CET6 Essay: Level of difficulty to meet the requirements: organizing the essay properly.	low	10	3.20	.63
	mid	12	2.75	.62
	high	11	3.55	.82

Note. SD=Standard deviation.

Table 13: Multiple comparisons between groups (Bonferroni): writing tasks

Dependent Variable	(I) Scorer Groups	(J) Scorer Groups	Mean Difference (I-J)	Std. Error	Sig.
1. IB-CET6 Summary: While reading the passage, I paid attention to the key information it contains.	low	Mid	.00	.33	1.00
		High	-1.14*	.34	.01
	mid	Low	.00	.33	1.00
		High	-1.14*	.32	.00
	high	Low	1.14*	.34	.01
		Mid	1.14*	.32	.00
2. IB-CET6 Summary: While reading the passage, I associated the passage with my previous knowledge.	low	Mid	.20	.47	1.00
		High	-0.94	.48	.18
	mid	Low	-0.20	.47	1.00
		High	-1.14	.46	.06
	high	Low	.94	.48	.18
		Mid	1.14	.46	.06
3. IB-CET6 Summary: After writing, I examined if my personal opinions had been included into the summary.	low	Mid	-1.07	.49	.12
		High	-1.31*	.50	.04
	mid	Low	1.07	.49	.12
		High	-0.24	.48	1.00
	high	Low	1.31*	.50	.04
		Mid	.24	.48	1.00
4. IB-CET6 Summary: After writing, I examined if my summary was coherent.	low	Mid	-0.63	.41	.41
		High	-1.16*	.42	.03
	mid	Low	.63	.41	.41
		High	-0.53	.40	.59
	high	Low	1.16*	.42	.03
		Mid	.53	.40	.59
5. IB-CET6 Summary: After writing, I refined my language.	low	Mid	-0.55	.43	.63
		High	-1.48*	.44	.01
	mid	Low	.55	.43	.63
		High	-0.93	.42	.10
	high	Low	1.48*	.44	.01
		Mid	.93	.42	.10
6. IB-CET6 Essay: Before writing, I planned the procedure and allocated time for each stage.	low	Mid	-0.78	.44	.27
		High	.40	.45	1.00
	mid	Low	.78	.44	.27
		High	1.18*	.42	.03
	high	Low	-0.40	.45	1.00
		Mid	-1.18*	.42	.03
7. IB-CET6 Essay: Factors affecting my performance: time allowed for writing.	low	Mid	-0.11	.37	1.00
		High	-1.14*	.38	.02
	mid	Low	.11	.37	1.00
		High	-1.03*	.35	.02
	high	Low	1.14*	.38	.02
		Mid	1.03*	.35	.02
8. IB-CET6 Essay: Level of difficulty to meet the requirements: organizing the essay properly.	low	Mid	.45	.30	.43
		High	-0.35	.31	.80
	mid	Low	-0.45	.30	.43
		High	-0.80*	.29	.03
	high	Low	.35	.31	.80
		Mid	.80*	.29	.03

Note. *p<.05.

To summarize, high-scorers, on the whole, were found to be more skilled in manipulating source materials and using metacognitive strategies than low scorers. The more advanced language users not only had a higher language proficiency level in pronunciation, vocabulary, and syntactic structure, but also tended to pay more attention to contextual clues that would facilitate their understanding of the materials, or help them organize the discourses.

3.3.2 Cognitive processing of independent and integrated tasks (t-tests)

T-tests elicited 24 questions with significant differences between the components of the two tests, indicating different cognitive processes elicited by tasks of a different nature. Below is a detailed review of the results.

Listening

Paired sample t-tests identified four questions with statistically significant differences between the two tests (Table 14). The integrated tasks in PTE Academic required that test takers use more *note-taking strategy* (MD=-0.58); test takers reported that they had more *background knowledge about the listening materials* of PTE Academic (MD=-0.58); test takers were more likely to be affected by *pronunciation* (MD=-0.49) and *gender of the speakers* (MD=-0.59) in PTE Academic.

Table 14: Mean differences of processes: listening tasks

Question	IB-CET6	PTE Academic	MD	Sig.	*d*
6. I took notes to facilitate my memorizing and understanding.	2.97	3.55	-0.58	.04	.51
8. I had background knowledge that facilitates my understanding.	2.39	2.97	-0.58	.00	.65
13. Influence: pronunciation of the speaker	3.09	3.58	-0.49	.03	.52
16. Influence: gender of the speaker (male/female)	2.16	2.75	-0.59	.01	.56

Notes. MD=Mean difference (IB-CET6–PTE Academic); $p<.05$ (two-tailed); d=Cohen's (1988) *d,* $d=.2$ (small), $d=.5$ (medium), $d=.8$ (large).

Speaking

Six questions with significant mean differences between the two tests were identified (Table 15), with *Q4: I focused on how I should organize my speaking* having a large effect size ($d=1.01$). Independent speaking tasks in IB-CET6 seemed to have put more "mental burdens" on test takers than integrated tasks: the candidates found it more difficult to *express their ideas accurately* (MD=.33), paid more attention to *content* (MD=.54) and *organization* (MD=.88), and were more influenced by the *topic* given (MD=.79). However, independent speaking tasks seemed to facilitate the use of *English* in executive processing (MD=.51) while *Chinese* was more frequently used in manipulating source materials in integrated tasks (MD=-0.75).

Table 15: Mean differences of processes: speaking tasks

Question	IB-CET6	PTE Academic	MD	Sig.	*d*
3. I focused on the content of my speaking.	3.88	3.34	.54	.00	.61
4. I focused on how I should organize my speaking.	3.73	2.85	.88	.00	1.01
7. I considered/wrote down my speech content in English.	3.09	2.58	.51	.04	.39
8. I considered/wrote down my speech content in Chinese.	2.67	3.42	-0.75	.01	.54
18. Influence: speech topic	3.88	3.09	.79	.01	.63
23. Difficulty: conveying the thoughts accurately	3.12	2.79	.33	.02	.41

Notes. MD=Mean difference (IB-CET6–PTE Academic); $p<.05$ (two-tailed); d=Cohen's (1988) *d*, $d=.2$ (small), $d=.5$ (medium), $d=.8$ (large).

Reading

Eight questions were identified with significant differences in mean scores (Table 16). The strategy of *reading the passage before reading and answering the questions* was significantly less used in IB-CET6 fast reading than in PTE Academic. In both IB-CET6 fast and in-depth reading, test takers felt that they had a better *control of time* (MD=.73) than in PTE Academic (MD=.69). In PTE Academic, they were more likely to *answer the questions that they were able to answer first* (MD=-0.46; -0.52). PTE Academic also seemed to have engaged test takers in more careful reading than IB-CET6: test takers were more likely to *grasp the main idea of each paragraph* (MD=-0.51), *read through every paragraph carefully* (MD=-1.00, *d*=1.00) and *read the passage word by word* (MD=.58).

Table 16: Mean differences of processes: reading tasks

Question	IB-CET6	PTE Academic	MD	Sig.	*d*
FR2. I read the passage first, then I read and answered the questions.	1.85	2.39	-0.54	.01	.50
FR4. I had a good control of time for all questions.	3.73	3.00	.73	.00	.71
FR6. I grasped the main idea of each paragraph.	2.73	3.24	-0.51	.02	.56
FR15. I answered the questions that I was able to answer first.	3.45	3.91	-0.46	.04	.39
IR4. I had a good control of time for all questions.	3.69	3.00	.69	.00	.72
IR5. I read through every paragraph carefully.	2.30	3.30	-1.00	.00	1.00
IR12. I did not read the passage word by word.	3.97	3.39	.58	.03	.52
IR15. I answered the questions that I was able to answer first.	3.39	3.91	-0.52	.04	.44

Notes. FR=Fast reading, IR=In-depth reading; MD=Mean difference (IB-CET6–PTE Academic); $p<.05$ (two-tailed); *d*=Cohen's (1988) *d*, *d*=.2 (small), *d*=.5 (medium), *d*=.8 (large).

Writing

Six questions were identified with significant differences in mean scores (Table 17). The three questions on summary writing were all related to the source material used in the task: IB-CET6 summary writing seemed to require test takers to pay more attention to *the key information in the reading passage* (MD=.55), whereas in PTE Academic test takers were more concerned with the *genre* (MD=-0.49) and *topic* of reading passages (MD=-0.49). In the essay writing task of both tests, test takers *seldom drafted their essays*, and this was even more likely to be the case with IB-CET6 (MD=-0.61). For the two factors affecting writing performance, test takers seemed to be more concerned with the *topic* (MD=.68) and *required length* of the essay (MD=.40) in IB-CET6 than in PTE Academic.

Table 17: Mean differences of processes: writing tasks

Question	IB-CET6	PTE Academic	MD	Sig.	*d*
SW4. I paid attention to the key information it contains.	3.88	3.33	.55	.01	.59
SW26. Influence: genre of the reading passage	3.33	3.82	-0.49	.01	.50
SW33. Influence: the extent to which I am interested in the topic	3.39	3.88	-0.49	.02	.52
EW7. I drafted the essay by writing something down.	2.18	2.79	-0.61	.03	.52
EW18. Influence: writing topic	4.06	3.38	.68	.00	.76
EW19. Influence: the required length of writing	3.52	3.12	.40	.04	.43

Notes. SW=Summary writing, EW=Essay writing; MD=Mean difference (IB-CET6–PTE Academic); $p<.05$ (two-tailed); d=Cohen's (1988) d, $d=.2$ (small), $d=.5$ (medium), $d=.8$ (large).

In summary, in tasks that measure test takers' receptive abilities, the differences mainly concern the distinctive features of spoken language in listening materials, and strategies and skills in reading tasks. In tasks that assess test takers' productive skills, i.e., speaking and writing, the identified differences mainly relate to the topic and executive processes of preparing. Most differences had a medium effect size ($.50 \leq d < .80$); the effect sizes of two questions, i.e., *I focused on how I should organize my speech before starting to speak* ($d=1.01$) in the speaking section and *I read through every paragraph carefully* ($d=1.00$) in the reading section, were large, suggesting that task type may seriously affect test takers' cognitive processes in these two aspects.

4. Discussion

4.1 Potential risks of "muddied measurement" in integrated tasks

A very high correlation ($r=.95$) was reported in Zheng and De-Jong (2011) between PTE Academic and TOEFL iBT, both employing predominantly tasks of an integrated nature. The slightly lower correlation between PTE Academic and IB-CET6 ($r=.84$) reported in this study could, therefore, be explained by the effect of task type on test takers' overall performance and/or score reporting. Interestingly, the correlation between PTE Academic and IB-CET6 is stronger than that between PTE Academic and IELTS ($r=.73$), as reported in the same study by Zheng and De-Jong (2011). A possible explanation for the lower correlation is that test delivery mode (paper-based or computer-based) may also have an effect on test takers' overall performance.

Results of the analyses of test score data at the component skill level, however, caused concern over the issue of "muddied measurement" with tasks of an integrated nature (Urquhart & Weir, 1998; Weir, 1990). On the one hand, the internal correlations between the four components of PTE Academic were higher than what we had expected. The listening and writing scores of PTE Academic, for example, were correlated at .90. Is such a high correlation a true reflection of the comparability between test takers' listening and writing abilities? If yes, then what is the purpose of reporting profile scores of listening and writing? Or, is it necessary for the test to retain both components? On the other hand, the relatively low correlation between the subscores of PTE Academic and IB-CET6 raised further doubts about the practice of profiling test takers' communicative skills, based on their performances on tasks of an integrated nature.

A further argument against profile score reporting is that integrated tasks engaging test takers in multi-modality language activities require both integrative manipulation of information and integrated processing of language skills (Anderson, 2009; Iwashita, 2008). The investigation of the cognitive processing involved in task completion in this study revealed that different executive processing strategies were adopted in integrated and independent tasks. *Note-taking*, for example, was reported to be a useful strategy for PTE Academic listening tasks, but not for IB-CET6 listening tasks. The executive process of *paying attention to key information in the reading text* for summary writing was reported to be a useful strategy by IB-CET6 test takers, and high-scorers were better at utilizing this strategy than low-scorers. In essay writing, instead, *time management* was found to be a key facet differentiating high-scorers from low-scorers. Independent speaking tasks seemed to facilitate the use of *English* in executive processing while *Chinese* was more frequently used in manipulating source materials in integrated tasks.

4.2 Suggestions for score reporting

It is suggested that the target (or primary) modality/skill of each integrated task be explicitly stated and appropriately weighed if the current practice of reporting the four communicative skills is to be retained. For example, the task *read aloud* is targeted more specifically at speaking. The reported score of this task should give more weight to speaking than reading. Similarly, score report of the task *summarize written text* should give more weight to writing than reading. Decisions on weighting are, nonetheless, difficult to make, and would perhaps be largely based on experience. However, given the increasing popularity of integrated tasks in language testing and assessment, further exploration of the way to report performance on integrated tasks will prove a worthy effort.

In addition to the total and subscores, PTE Academic also reports scores of six enabling skills, including grammar, oral fluency, pronunciation, spelling, vocabulary, and written discourse. The information of these enabling skills is diagnostically very useful for test takers to improve their language abilities, but not so much for test users, who are much more concerned about whether the candidate is able to use the language to fulfill real-world language tasks. It is, therefore, suggested that tests relying heavily on integrated tasks should investigate the possibility of profile score reporting at either task level or module level. Reporting task-level scores may have practical difficulties because a test usually has quite a number of tasks (e.g., 20 tasks in PTE Academic). Reporting module-level scores may prove a more practical and useful solution. "Module" here refers to the way language skills are assessed, independently or in combinations. In this sense, PTE Academic has altogether nine modules: the four independent modules of *listening*, *speaking*, *reading*, and *writing*, and five integrated modules of *reading and speaking*, *listening and speaking*, *reading and writing*, *listening and writing*, and *listening and reading*.

Admittedly, such changes in the way test scores are profiled and reported may have practical implications for score equation, and more importantly, theoretical implications for score interpretation. However, if the purpose of integrated tasks is to better simulate real-life communicative activities, and language skills are indeed integrated in the real world, would it be equally meaningful to report performances on multi-modality tasks such as "listening and writing", "reading and speaking", apart from scores of listening, reading, speaking, and writing?

4.3 Improvement of test fairness in integrated tasks

According to the results of t-tests, it is found that in the present study, one

significant contributor to the differences in cognitive validity between independent and integrated tasks lies in the scope of *content knowledge*. Namely, unlike integrated tasks that feature the *manipulation of external knowledge/source materials*, in independent tasks, the test takers' performances appear to be highly associated with their *internal knowledge* (e.g., *Q3*, *Q4*, and *Q18* in the speaking section; *Q18* in essay writing).

In independent tasks measuring test takers' productive skills such as speaking and writing, the long-term memory that includes topical knowledge and linguistic knowledge, according to the information processing theory in L2 speech performance (e.g., Ashcraft, 1994; Atkinson & Shiffrin, 1968) and the model of writing process (Hayes & Flower, 1980), is an influential factor for the speaker's/writer's performance. In writing tasks, particularly, such long-term memory and internal content knowledge permeate the whole writing process including "planning, drafting, revising and editing" (Hyland, 2002:25). However, from the perspective of test fairness, topical knowledge is regarded as one of the construct-irrelevant factors that may threaten test validity (Kunnan, 2000: 3). In this sense, integrated tasks can to a large extent improve test fairness in that the input information saves test takers' efforts to generate the topical content from long-term memory and organize the logic sequence of a discourse (Plakans, 2008). Since topical knowledge or content for task completion is made accessible in integrated tasks, chances become low that test validity would be reduced by construct-irrelevant factors such as lacking topical knowledge.

4.4 Limitations of the study

The results of the study, nevertheless, should be interpreted with caution due to the limitations in the sample size and research context. Though great efforts were made to invite test takers to participate in this study, we only managed to get 43 candidates for the main study, of which 10 did not produce valid data. The small sample size made it impossible for us to conduct factor analysis, as originally proposed, and rendered the conclusions of the study more tentative. The main reason for the difficulty in recruiting participants is that targeted candidates of the two tests simply do not overlap. PTE Academic is taken by university graduates applying to higher education institutions abroad, whereas IB-CET6 is given to university students during their first and second years as an indication of whether they have met College English Curriculum Requirements in China. In other words, there is no motivation on the part of the candidate to take both tests. The other major limitation is that the study was conducted in only one type of context: English language learning at the tertiary level in China. Although none of the participants had the experience of taking IB-CET6 before this study, university students/graduates in China are nonetheless more familiar with locally developed English language tests, which usually adopt independent tasks. This may also have affected the participants' perceptions of task difficulty and cognitive processes involved in task completion. Therefore, to gain a fuller picture and a deeper understanding of skill integration in language assessment, further studies involving more test takers from different testing contexts are needed for an in-depth analysis of test performance, as well as processes involved in task completion.

References

Ashcraft, M. H. (1994). Human memory and cognition. New York: Harper Collins.

Atkinson, R. C. & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spencer (Ed.), The psychology of learning and motivation: Advances in research and theory (Vol.2, pp.89-195). New York: Academic Press.

Cohen, J. (1988). Statistical power and analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum.

Friend, R. (2001). Effects of strategy instruction on summary writing of college students. Contemporary Educational Psychology, 26(1), 3-24.

Garner, R. (1982). Verbal-report data on reading strategies. Journal of Literacy Research, 14(2), 159-167.

Hayes, J. R. & Flower, L. S. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), Cognitive processes in writing (pp.31-50). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hyland, K. (2002). Teaching and researching writing. London: Longman.

Kennedy, M. L. (1985). The composing process of college students writing from sources. Written Communication, 2(4), 434-456.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium (pp.1-15). UK: Cambridge University Press.

Larson-Hall, J. (2010). A guide to doing statistics in second language research using SPSS. New York and London: Routledge.

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. Assessing Writing, 13(2), 111-129.

Urquhart, S. & Weir, C. J. (1998). Reading in a second language: Process, product and practice. London: Longman.

Weir, C. J. (2005). Language testing and validation: An evidence-based approach. New York: Palgrave Macmillan.

Yang, L. & Shi, L. (2003). Exploring six MBA students' summary writing by introspection. Journal of English for Academic Purpose, 2(3), 165-192.

Zahedi, K. & Shamsaee, S. (2012). Viability of construct validity of the speaking modules of international language examinations (IELTS vs. TOEFL iBT): Evidence from Iranian test takers. Educational Assessment, Evaluation and Accountability, 24(3), 263-277.

Zheng, Y. & De-Jong, J. (2011). Research note: Establishing construct and concurrent validity of Pearson Test of English Academic. http://pearsonpte.com/research/Documents/RN_EstablishingConstructAndConcurrentValidityOfPTEAcademic_2011.pdf.