Research Note: Demonstrating cognitive validity and face validity of PTE Academic Writing Items *Summarize Written Text* and *Write Essay*)

Written by
Sathena H. C. Chan, MA University of Essex, UK

1. Introduction

This paper is based on a Master's thesis submitted to the University of Essex, in September 2010. The research was funded by Pearson as part of the 2010 Pearson External Research Program and was supervised by Tony Lilley, Department of Language and Linguistics at the University of Essex.

This study examines the cognitive validity of two item types of the Writing Section of the PTE Academic test – **Summarize Written Text** and **Write Essay** - within Weir's (2005) socio-cognitive framework for test validation. The study focuses on cognitive validity by investigating and comparing the cognitive processes of a group of ESL test takers undertaking **Summarize Written Text** (an integrated writing item) and **Write Essay** (an independent writing item). Cognitive validity is a 'measure of how closely it [a writing task] represents the cognitive processing involved in writing contexts beyond the test itself' (Shaw and Weir, 2007:34). In addition, the study investigates test takers' opinions regarding the two different writing item types: independent and integrated. Test takers' scores on both items are compared to investigate if the two performances correlate.

The study uses screen capture technique to record test takers' successive writing processes on both items, followed by retrospective stimulated recalls. The findings demonstrate that **Summarize Written Text** and **Write Essay** engage different cognitive processes that are essential in academic writing contexts. In particular, macro-planning and discourse synthesis processes such as selecting relevant ideas from source text are elicited by the **Summarize Written Text** item whereas processes in micro-planning, monitoring and revising at low levels are activated on the **Write Essay** item. In terms of test performances, the results show that test takers in this study performed significantly better on **Write Essay** than on **Summarize Written Text**.

Concerning face validity, while most test takers hold a positive attitude towards **Summarize Written Text** when used for teaching and learning purpose, they prefer **Write Essay** to be used in a testing context. However, a careful examination of the reasons provided by the test takers suggests that their preferences reflect the perceived easiness rather than appropriateness of the item types.

2. Literature review

2.1 Independent and integrated item types

Independent writing-only items have been used extensively in large-scale academic writing tests (Weigle, 2002; Shaw and Weir, 2007). However, research on predominant academic writing tasks argues that impromptu argumentative essays represent only one type of writing tasks among many others that are required in real-life academic writing contexts (Carson, 2001; Cumming et al., 2005; Horowitz, 1986). Therefore, there is a need to diversify the types of writing items which are employed to measure academic abilities in order to improve the construct validity of a writing test and to bring more positive washback effects of the test (Hamp-Lyons and Kroll, 1997). PTE Academic is one of the large-scale academic English proficiency tests which include both independent and integrated writing items.

Academic writing is rarely done in isolation, but is always done in response to source texts Hamp-Lyons & Kroll, 1997; Weigle, 2004). It is generally believed that reading-to-write item types are more authentic than writing-only item types (Read, 1990; Weigle, 2004). However, it is important to be aware of the fact that the use of reading-to-write writing tasks does not offer a simple solution to the problems found with independent writing tasks (Plakans, 2008). Fundamental issues concerning the nature of integrated item types remain unresolved for many stakeholders. For instance, some researchers have questioned the possibility of a 'muddied measurement' (Weir, 2005: 101) due to the confusing effects of reading and writing abilities on the reading-to-write performance (Alderson, Clapham, & Wall, 1995). Therefore, evidence of what constructs are measured by integrated and independent items is needed.

Messick (1995) proposes that validity should be a multifaceted framework resulting from an integration of content, criteria and consequences. More recently, Weir's socio-cognitive approach (2005) demonstrates how such a validation framework can be achieved. He argues that the construct of any writing test items has to be demonstrated by three interrelated dimensions of evidence: context, cognitive processing and scoring. This study concentrates on demonstrating only the cognitive validity of the two PTE Academic item types. As Weir stresses, the evaluation of cognitive validity involves collecting both a priori evidence on the cognitive processes elicited by the item and a posteriori evidence involving statistical analysis of scores following test administration. The present study collected both pieces of evidence from the two item types **Summarize Written Text** and **Write Essay**.

2.2 Research comparing independent and integrated writing performances

As the use of integrated writing items (reading-to-write in particular) in assessing academic writing is increasing (Plakans, 2008), a number of studies have examined similarities and differences between independent and integrated writing performances. Most findings show that writers' performances on reading-to-write items do not always correlate to their performances on writing-only items (Esmaeili, 2002; Plakans, 2008). For example, Yu (2008) investigated integrated writing performances on summary tasks of 157 Chinese writers in two languages. His findings show that the participants' independent writing abilities (TOEFL writing-only scores) did not correlate significantly with their summary writing performances. In addition, research indicates that resulting text features produced from reading-to-write and writing-only items are different (Cumming et al, 2005). Asencion-Delaney (2008) argues in her study on the reading-to-write construct that reading-to-write, on the one hand, is a unique construct which is different from either the reading comprehension or writing-only construct. On the other hand, reading-to-write is a dynamic construct which consists of many dimensions depending on the specific item type used in the test.

2.3 Writing models

A number of researchers have attempted to propose a model of writing that accounts for the cognitive processes involved in writing (mainly L1 writing) and components that interact with these processes. Hayes and Flower's (1983) model challenges the perception that writing is a linear process. They propose that writing is indeed an extended problem-solving exercise which involves multiple recursions of *planning, translating* and *reviewing*. Bereiter and Scardamalia (1987), on the other hand, examine the differences of writing processes employed by immature and mature writers. Their model proposes a distinction between *knowledge telling* and *knowledge transforming* to differentiate writing as a rather linear memory retrieval process from writing as a recursive problem-solving process. Grabe and Kaplan's (1996) model, one of the few models of L2 writing, emphasizes the important roles of *internal goal setting, metacognitive awareness* and *monitoring* in L2 writing. Other researchers investigate how writers read texts for the purpose of academic writing. For example, Spivey's (1984) discourse synthesis model proposes that writers construct meaning from reading for a new text through *organizing* structure, *selecting* relevant content and *connecting* content with own knowledge. The present study investigated whether and to what extent the above mentioned writing processes are elicited by **Summarize Written Text** and **Write Essay**.

2.4 Research comparing writing-only and reading-to-write processes

Previous literature which compares writing processes on different writing-only (mainly argumentative essay) and reading-to-write items (such as argumentative essay, summary writing, journals comparison, etc.) suggests that writing processes are affected by the nature of an item type.

In general terms, writing-only items tend to elicit longer processes in goal-setting, planning and organizing ideas before writing however once writers have started to write (i.e. to transfer ideas into linguistic forms), they tend to write in a rather linear manner and do less editing while writing (Plakans, 2008; Severinson-Eklundh & Kollberg, 2003). Nevertheless, writers tend to do more rereading of their writing (Plakans, 2008)

On the other hand, while reading-to-write items seem to elicit shorter goal-setting and planning before writing (Plakans, 2008), writers engage in discourse synthesis processes such as reading source text, accepting or rejecting viewpoints in reading, re-ordering information, adjusting arguments, etc. (Esmaeili, 2002). Writers also seem to do more monitoring and revising at an advanced level (Severinson Eklundh and Kollberg, 2003). The overall reading-to-write process also tends to be more recursive than the writing-only process. However, since the reading-to-write and writing-only tasks used in previous studies vary and the scope of these studies is limited, caution is needed when we interpret the results regarding the overall difference between reading-to-write and writing-only processes.

The previous studies provide a general picture of the writing-only and reading-to-write process. To the author's knowledge, there has not yet been an attempt to compare writing processes elicited by two item types under test conditions. The present study thus aims to address three research questions:

Demonstrating cognitive validity and face validity of PTE Academic writing items *Summarize Written Text* and *Write Essay*

3

Research Questions

(1) Do test takers employ composing processes differently on two PTE Academic item types: **Summarize Written Text** and **Write Essay**?

(2) What are the test takers' opinions concerning **Summarize Written Text** and **Write Essay** and how effectively do they believe each of these item types serve as a teaching task or a testing task? What are their reasons for holding these views?

(3) How do test takers perform on **Summarize Written Text** and **Write Essay**? Is there any correlation between the two performances?

3. Research Methods

A mixed method approach was used in this study (see Table 1). Ten performances on **Write Essay** and **Summarize Written Text** were recorded in a non-intrusive manner by using Camtasia studio, a screen capture software program published by TechSmith. Lengths of different writing processes were recorded and then analysed. The test takers' writing processes were also observed by the researcher. Immediately after each writer completed the items, the screen video was used in a stimulated recall of his/her writing processes. Questions were asked to prompt the test taker to describe their writing process while they were watching the video. After the stimulated recall, each test taker was interviewed about his/her perception of the two PTE items **Summarize Written Text** and **Write Essay** and how effectively they believed each of these item types reflected their writing ability.

Table 1: Data collection methods		
Stages	**Instrument**	**RQs to be addressed**
While writing	**Summarize Written Text Write Essay**	RQ1: process RQ2: face validity RQ3: performance
	Screen video recording	RQ1: process
	Observation	RQ1: process
After writing	Stimulated recall on writing process	RQ1: process
	Interview on opinions	RQ2: face validity
	Score analysis	RQ3: performance

3.1 Participants

Ten participants took part in this study. They were all international postgraduate students at the University of Essex. These students came from four different L1 backgrounds: Arabic (n=5), Chinese (n=3), Japanese (n=1) and Malay (n=1). They were recruited from the same department, i.e. the Department of Language and Linguistics, to minimize potential topic effect on individual participants. In addition, the participants were selected on basis of their second language writing proficiency, i.e. IELTS writing scores 6 to 7.

3.2 Test items

PTE Academic is an international computer-based academic English test that comprises twenty item types. Each item type assesses the listening, reading, speaking or writing ability of test takers separately or in an integrative manner. For more details regarding PTE Academic, please see Pearson (2009). The reading-to-write and writing-only tasks used in this study were taken from the writing section of PTE Academic Practice Test, which measures test takers' ability to produce written English in academic settings (Pearson, 2009:6) .

The reading-to-write task consists of one item of the item type **Summarize Written Text**. The task requires test takers to write a one-sentence summary of a passage after reading a text in ten minutes. The test taker has to complete the task within ten minutes. The topic of the item used in this study was 'democracy'.

The writing-only task consists of one item of the item type **Write Essay**. The task requires test takers to write an essay of 200 to 300 words about a given topic. The test taker has twenty minutes to complete the task. The topic of this particular item was 'advertising'.

Two items of each item type were piloted with three students with a similar background as the participants in the main study. The students completed all four items and commented on each item, input text difficulties as well as their familiarity of the topics. One item was selected from each item type based on the students' comments and their performances on the items.

3.3 Scoring

In most of the previous process studies, scripts were not scored due to the interference in the writing process. The use of a non-intrusive method in this study allows writers to produce the scripts in authentic testing conditions, e.g. under timed-condition. The scripts were scored by Pearson automated scoring technology following two specific scoring rubrics for each task type. According to Pearson (2009), *Summarize Written Text* is scored on five traits while *Write Essay* is scored on seven traits. Both tasks are scored based on the quality of the content, the fulfillment of the formal requirement, the accuracy of grammar, the range and appropriateness of vocabulary use and spelling. In addition, *Write Essay* is scored on the development, structure and coherence of ideas, and general linguistic range.

4. Data analysis

4.1 Writing process types analysis

Ten screen videos of *Summarize Written Text* and *Write Essay* were analyzed in intervals of one second. Most previous studies have analyzed writing process in a range of one to six seconds (Kowal & O'Connell, 1987). In total, there were 6000 seconds of *Summarize Written Text* data (600 seconds each *10) and 12000 seconds of *Write Essay* data (1200 seconds each *10). The data were analyzed according to the following thirteen process types: Reading rubrics/source text and planning, Writing, Global editing word level 1, Global editing word level 2, Global editing sentence Level, Local editing word level 1, Local editing word level 2, Local editing sentence level, Pausing, Moving cursor, Checking interface/remaining time/word count, Cursor error and Unused time. Some categories, for example, reading rubrics/source text (Plakans, 2008), pausing (Bosher, 1998), and global and local editing (Severinson-Eklundh & Kollberg, 2003), were pre-defined by following previous studies on writing process. Other categories, such as cursor error and unused time, were added by the researcher after analyzing the pilot data. See Table 2 for the detailed definition of each process type. Ten percent of the data was co-rated by an employed PhD student. The initial agreement rate was 91%. Most of the disagreement concerned the editing performed to undo typing mistakes caused by 'cursor error'. After discussing the differences and agreeing on a standard interpretation, the agreement rate was 96%.

4.2 Statistical analysis

After identifying all instances according to the thirteen process types from the ten *Summarize Written Text* and ten *Write Essay* screen videos, the length of time (in seconds) that the test takers spent on each process type and the frequency of occurrence of each process type across the two items were computed and then analyzed statistically. Due to the small sample size, the Wilcoxon Test for paired data was used to test the differences. Apart from the amount of time spent on each process type and the frequency of their occurrence, the order of how the test takers employed each process type was also studied quantitatively.

Table 2: Definition of process types	
Process Type	**Definition**
Reading and planning	This type occurs when the writer reads the rubrics (and source text) and plans before writing.
Writing	This type occurs when the writer composes the text.
Global editing	The writer interrupts the text production to make one or a sequence of revisions at different locations in the text written previously. The revisions may or may not be semantically related to one another. After the last revision in the sequence, the writer either resumes writing at the position of the interruption or pauses or ends the writing session.
Local editing	Immediate revisions are made at one cursor location, where the writer is currently producing text. This type occurs, for instance, when the writer is trying out different words in one place in the text to find the right way to express something, i.e. deleting and inserting repeatedly at the same position.
Word level 1	Revisions at word level regarding grammatical accuracy, e.g. capital letter, plural/singular, spelling mistake, typo, etc.
Word level 2	Revisions at word level regarding meaning, e.g. replacement by another word, etc.
Sentence level	Revisions at sentence level. This type also includes the writer replacing one word with another if the change affects the sentence structure.

Demonstrating cognitive validity and face validity of PTE Academic writing items *Summarize Written Text* and *Write Essay*

5

Pausing	This type occurs when the writer pauses after the writing begins. This may include processes such as rereading prompt/source text/ monitoring/ organizing ideas/ reading own text, etc.
Moving cursor	This type occurs when the writer moves the cursor. This may include processes such as rereading prompt/source text/ monitoring/ organizing ideas/ reading own text, etc.
Cursor error	This refers to the cursor being moved accidentally away from the current writing location without the writer's intention. It occurs for instance when the writer mistakenly touches the mouse pad.
Checking time/word count/interface	This type occurs when the writer explores the interface or checks the remaining time or the word count.

4.3 Stimulated recall data

All stimulated recall data were transcribed[1] and then analyzed in regard to what aspects of their writing the test takers had been attending to while completing the tasks. The coding scheme used was developed by Cumming et al. (1989) and adapted by Bosher (1996) (see Table 3 for the coding scheme). Ten percent of the data was co-rated by an employed PhD students and the agreement rate was 76%. The major disagreement was due to an apparent overlap between gist, discourse organization and procedure. When writers described what they were doing (procedure), they often mentioned their ideas (gist) and the structure of their writing (discourse organization) at the same time. One possible solution was to use a more detailed coding scheme, but the present coding scheme was retained for the following reasons. Firstly, the stimulated data was not primary data in this study but a supplement to the screen video data. Secondly, this scheme was used in the previous reading-to-write studies (e.g. Bosher, 1998). After revisiting the coding scheme with the co-rater a few times, it was decided that if the writers mentioned content or structure while they were describing their procedure, that part of transcriptions would be coded under 'procedure'. But in order to focus the analysis on the composing processes, the 'procedure' category was further divided into different sub-procedures, e.g. reading instruction, reading prompt, reading own text, etc.

Table 3: Coding scheme	
Attention to aspects of writing	
Gist	Substantive content of the writing – the writer's thoughts or ideas
Discourse Organization	Organization of written discourse, its structure beyond the level of the clause
Intention	Overall purpose of the text or a portion of the text
Language Use	Use of English as a linguistic code
Procedure	Reference to procedural issues from text generation to difficulties with fluency

5. Results and Discussion

5.1 Overall writing process

Due to the difference in the actual time allowed for *Summarize Written Text* (i.e. 10 minutes) and *Write Essay* (i.e. 20 minutes), the results presented have been standardized to show the percentage of the time spent on each process type. For the purpose of revealing a global picture of the writing process, the thirteen process types were collapsed into six: Reading rubrics/source text and Planning, Writing, Global editing, Local editing, Pausing and Unused time. Figure 1 shows the overall patterns in time allotted to different writing processes across the *Summarize Written Text* and *Write Essay* items. The qualitative results generated from the screen videos supplemented by the qualitative stimulated recall data are discussed in details in the following sections.

---

[1] The transcription was done by a student who was paid out of the research funding. The researcher went through all transcriptions to ensure their quality.
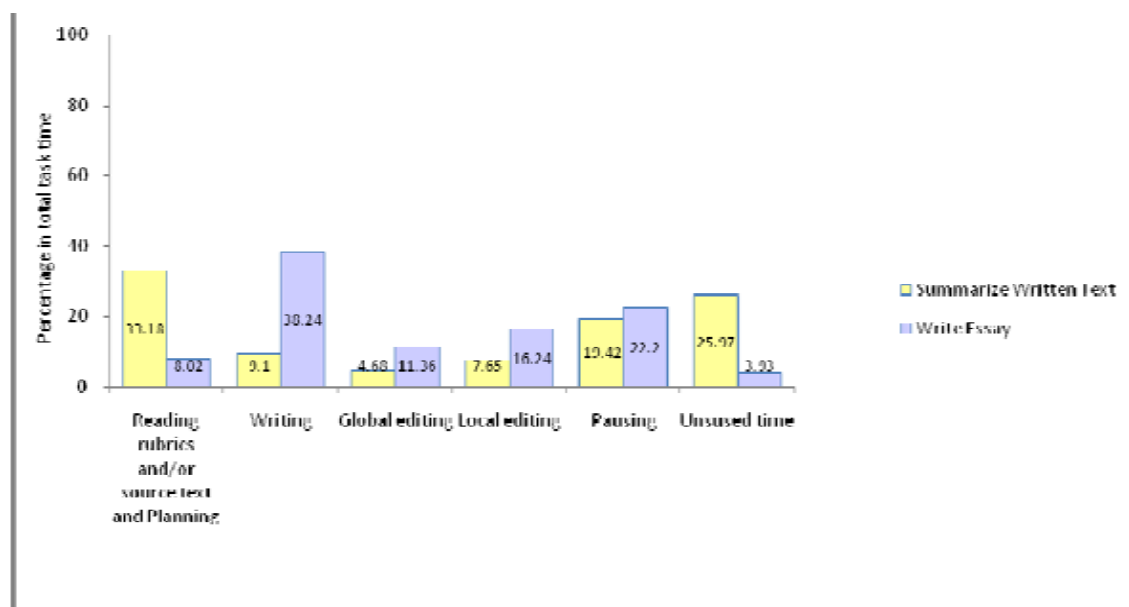
Figure 1: A comparison of process between *Summarize Written Text* and *Write Essay*

5.1.1 The preparing-to-write stage

The preparing-to-write stage which involves reading and planning reveals some differences between the two item types. Test takers spent on average 33.18% of the total time allowance on reading and planning before they started to write on the *Summarize Written Text* item while they spent just 8.02% on this stage on the *Write Essay* item. It is perhaps not surprising that the average time allotted to this stage on the *Summarize Written Text* item was four times (sig.=.005, p<0.5) more than that on the *Write Essay* item as test takers were required to read the source text in the former. However, the stimulated recall data show that the cognitive processes employed when responding to the items were distinctive. Plakans (2008) proposed a reading-to-write model based on ten postgraduates think-aloud protocols on a reading-to-write task (an argumentative essay writing task). Her model was adopted and modified to demonstrate the cognitive processes in which test takers engaged on the preparing-to-write stage across the *Summarize Written Text* and *Write Essay* items. See Figure 2a and 2b for the tentative models[2].

On the *Summarize Written Text* items, most test takers followed an overall pattern with some individual differences: most test takers quickly read the prompt once to understand the task and set their reading and writing goals, i.e. to identify key ideas from the source text and summarize them. The test takers tended to move rapidly from task interpretation and goal setting to reading the source text. All test takers read the source text more than once but the exact number varied. After grasping the general idea during the first reading, the test takers identified key ideas which they believed were relevant to the task in the selected parts of the source text. After that, test takers engaged in processes of selecting and organizing ideas for their writing. According to the test takers, detailed content planning was conducted before they started writing. It seems that task takers engaged in brief task interpretation but careful planning processes by reading source text, selecting ideas, connecting ideas to the task, organizing ideas.

In Plakans' (2008) study, only experienced and motivated writers interacted with the source text before they produced the essay on the reading-to-write task. However, the *Summarize Written Text* item appeared to successfully induce all test takers in this study to interact with the source text. Even the least motivated test taker who mentioned that her purpose was 'to finish' (P10[3]) explained how she selected and organized ideas from the source text. There seemed to be individual differences in how test takers interacted with the source text. For instance, some tried to 'capture what each sentence means' (P6), some tried to 'pick up some important sentences' (P5) or 'focus on the first sentences' (P7) whereas some chose to 'rely on important words and vocabulary to know the meaning of the message' (P3). The effectiveness of these strategies is yet another issue but such enforced interaction may well be seen as an advantage of a summary test task because the ability to interact with the source text is regarded as part of academic writing ability (Campbell, 1990).

---

[2] The models on the preparing-to-write stage were generated by the screen video data, supplemented by the test takers' explanations gathered from the stimulated recall.

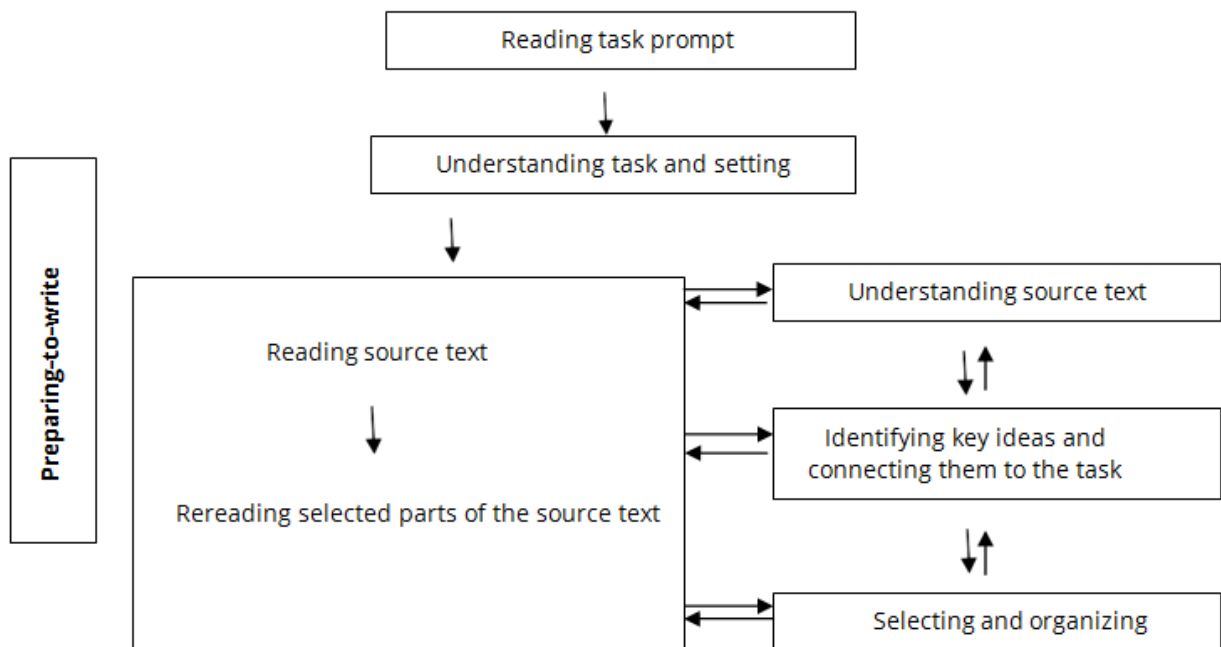[3] Participants were given a reference number from P1 to P10.

Demonstrating cognitive validity and face validity of PTE Academic writing items *Summarize Written Text* and *Write Essay*

7

Figure 2a: Preparing-to-write stage on the **Summarize Written Text** item

On the other hand, the **Write Essay** item seemed to elicit a comparatively longer process in task interpretation but a more mechanical planning process. All test takers read the whole item prompt a few times in order to 'understand the points of the question' (P3) and to 'find out the topic' (P3). As most test takers set their goal to 'write an essay' (P10), the item did not seem to provide test takers 'a reason for completing the task that goes beyond a ritual display of knowledge for assessment (Shaw and Weir, 2007:71). After task interpretation, most writers mentioned that they tried to identify key words from the prompt and link them to their personal knowledge at the same time. For example, one writer explained, 'I was thinking of my own experiences about what it means to vote for someone' (P8). But writers who were not familiar with the topic struggled to progress from positioning themselves with the topic to generating writing content. For example, one expressed his frustration: 'I'm not really familiar, and not care about these principles. I haven't thought about this question, so it's hard to create ideas. I didn't know what to write' (P10). This is perhaps why most test takers moved on to plan the structure instead of the content of the essay. One test taker recalled that 'for an essay, you have to write introduction and conclusion and some ideas in the middle' (P4). This kind of mechanical planning based on test takers' metacognitive knowledge of genre occurred often under timed testing conditions, especially for less proficient writers (Cumming et al, 2005).
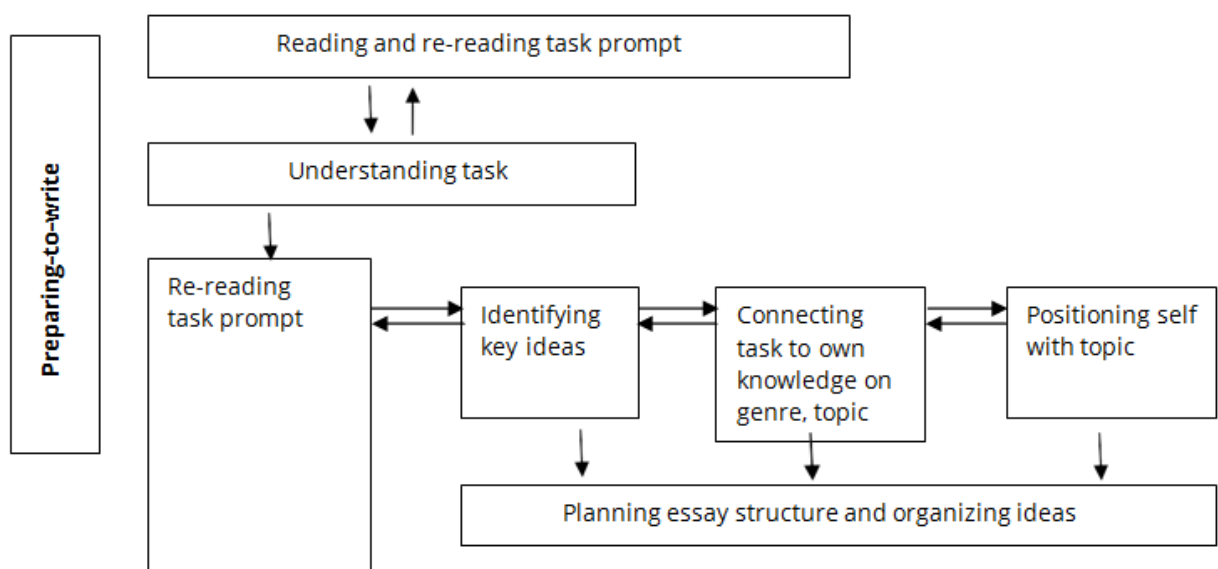

Figure 2b: Preparing-to-write stage on the **Write Essay** item

The preparing-to-write stage proposed in Plakans' (2008) model of composing process for reading-to-write tasks "was fairly linear, and writers moved step-by-step through preparing" (p.117). However, the findings generated from screen videos and stimulated recalls in this study show that test takers actually engaged in different patterns of circular and overlapping processes on such preparing-to-write stage when responding to the different item types.

5.1.2 Writing

After the preparing-to-write stage, test takers entered the writing stage in which they converted ideas to linguistic forms. There was a large difference, about 30%, (sig.= .005, p<0.5) in the total task time in writing (excluding the editing process) between the two item types. On average, test takers spent 9.10% of the time on writing on the **Summarize Written Text** item but 38.24% on the **Write Essay** item. This striking difference is not unexpected because the word requirement of the former task (i.e. one sentence) is much lower than that of the latter (i.e. 200-300). However, it is worth noticing that this result is parallel to Plakans' (2008) study in which the word requirement for both tasks was the same. It is also interesting to reveal that actual writing time was proportionally low compared to other processes, especially on the **Summarize Written Text** item. This leads to a further question about the exact roles other processes, such as monitoring and editing, play in the production of a text on either item types. The findings seem to support Grabe and Kaplan's (1996) emphasis on the role of internal goal setting, metacognitive awareness and monitoring in L2 writing.

5.1.3 Pausing behavior

Another major activity occurring during the writing stage was - pausing. The percentage of the time test takers paused from their composing process on the **Write Essay** item (22.2%) was slightly higher than that on the **Summarize Written Text** item (19.42%) but the difference is not significant. Pauses have been studied to reveal writers cognitive activities during writing. Matsuhashi (1982) believes that writers are likely to engage in cognitive planning and decision making behavior during pauses. In order to understand more about the test takers' pausing behavior in this study, further analysis[4] was performed and the results are summarized in Table 4 below. Despite the fact that the total pause time on the **Summarize Written Text** item was proportionally shorter, the mean pause length was actually slightly longer than that on the **Write Essay** item. The stimulated recall data reveals that the test takers paused for different reasons across the two items

Table 4: Descriptive statistics of pausing behavior		
Pausing behavior	Summarize Written Text	Write Essay
Total task time	600 sec	1200 sec
No. of pauses on average	7.5	22.9
Total pause time on average	84.2 sec	249.1 sec
Mean Pause Length on average	11.23 sec	10.88 sec
Percentage of total pause time on average	14.03 %	20.76 %

On the **Summarize Written Text** item, one major activity the test takers did during their pauses was to reread the source text. Their purpose in doing this might be to check their understanding of the source text (see Transcript excerpt 1a below); to search for words or phrases that they could use in their writing (1b & 1c); or to get more ideas for their writing (1d). Another activity the test takers did was to reread their own text. Some evaluated their own text from the readers' perspective (1e & 1f) while some did so from the rater's perspective (1g).

Transcript excerpt 1:
1a) 'Just to make sure that I understand it [the source text] pretty well.' (P1)
1b) 'I just want to include some words from the passage. I was finding the words.' (P10)
1c) 'I tried to copy the important part from the passage.' (P6)
1d) 'I reread the sentences because when I write I need more idea so I looked into the sentences and then write again.' (P3)
1e) 'I was thinking about the impact of the sentence. I wanted to make sure the reader is impressed by the first word.' (P2)
1f) 'When you are performer you make sure that what you perform will be received by the audience.' (P4)
1g) 'I was struggling not to use the same words from the passage because it is not allowed.' (P2)

---

[4] This analysis excluded data on moving cursor and checking interface/remaining time/word count.

Demonstrating cognitive validity and face validity of PTE Academic writing items *Summarize Written Text* and *Write Essay*

9

On the **Write Essay** item, the major reason why the test takers paused was to generate content by reading own text (see Transcript excerpt 2a below) or by recalling personal experiences from memory (2b). It was also common for the writers to try to elaborate the main idea they identified earlier (2c) or to recall vocabulary (2d). Besides, many test takers paused to check the grammar accuracy and the use of vocabulary (2e & 2f). It appears that the test takers were concerned with accuracy more on this item than they were on the **Summarize Written Text** item. A few writers were also concerned about the coherence of their text (2g), but the analysis of the screen videos does not show much evidence of corresponding high-level editing. Finally, some test takers paused due to writing blocks (2h).

Transcript excerpt 2:

2a) 'If I have the idea I can just continue but if I don't know what to write next, so I have to go back to see what I can write.' (P10)
2b) 'I am recalling my own experiences. I remember the voting just happened.' (P1)
2c) 'I tried to define democracy but I don't know how to explain ideas.' (P10)
2d) 'I was recalling the words I want.' (P8)
2e) 'I never go back reading the whole thing because I have to check everything grammatically, semantically before I go to the next one.' (P7)
2f) 'to make sure every sentence is very well organized and the grammar thing is good, vocabulary is good.' (P4)
2g) 'I don't want my points to be in segments I want the whole idea to link.' (P2)
2h) 'At this point I just cannot come with other good examples to support that concept of voting so I just gave it up.' (P4)

5.1.4   Global editing and the linearity of text production

The linearity of text production is another difference found between the writing process on the **Summarize Written Text** and **Write Essay** items. According to Severinson-Eklundh and Kollberg (2003), the writing process is linear if the text is produced in the exact order of its final presentation, whereas non-linear writing involves a lot of editing during writing, especially the global editing processes. As shown in Figure 1, test takers in this study did more global editing on the **Write Essay** item (11.36% of the total task time) than on the **Summarize Written Text** item (4.68%). However, given the large difference in the word requirement, i.e. one sentence for the former and 200-300 words for the latter, the difference in the amount of global editing was not significant (sig.=.059, p>.05). Another issue is when the test takers performed global editing. The screen videos reveal that global editing occurred late on the **Summarize Written Text** item. Most test takers edited their text globally during the last 30% of the time, while they edited globally throughout the **Write Essay** item.

It appears that there was no significant difference in the degree of linearity between the **Summarize Written Text** and **Write Essay** items but the patterns of linearity were different. The **Summarize Written Text** item was composed in a largely linear way until the last one-third of the production, when global editing occurred. This is different from the approaches identified in Severinson-Eklundh and Kollberg's (2003) study in which the production of the reading-to-write comparison task was non-linear and that of the reading-to-write argumentative task was of a mixed pattern. But the findings of different writing patterns across different reading-to-write tasks support Asencion-Delaney's (2008) argument that 'reading-to-write' is a collective construct which consists of different dimensions, dependent upon the specific task types. For the **Write Essay** item, although the test takers performed global editing throughout the task, 64% of the global edits was at word level. This seems to be parallel to the top-down planning model identified in Severinson-Eklundh and Kollberg's (2003) writing-only data. In other words, text takers tended to follow the original structure of the text throughout the composing process.

5.1.5   Local editing

Local editing, which is when writers made immediate editing at the current cursor location, was performed more often on the **Write Essay** item (16.24%) than on the **Summarize Written Text** item (7.65%) (see Figure 1). The difference is significant (sig.=.013, p<.05). Such processes may indicate when the writers make instantaneous corrections or try out different words (Severinson-Eklundh & Kollberg, 2003). Apart from the fact that test takers only needed to compose one sentence on the Summarize Written Test item, less local editing on this item might also be a result of test takers' careful planning during the preparing-to-write stage. On the **Write Essay** item, most test takers, according to the accounts they gave while watching the videos, were trying to generate content by recalling personal experiences while writing. This may explain why they needed to make more local edits. The stimulated recalls also reveal that test takers were very much concerned with accuracy while composing the **Write Essay** item.

5.1.6 Unused time

Another large difference (sig.=.028, p<0.5) across the two item types was the percentage of unused time. On the **Summarize Written Text** item, on average 25.97% of the total task time was not used whereas on average only 3.93% of the **Write Essay** item time remained unused. Seven test takers submitted their text before the allocated time was up for the **Summarize Written Text** item whereas only two writers did so for the **Write Essay** item. The correlation between the amount of unused time and the actual performance on the **Summarize Written Text** score was explored (see Figure 3). Even though the correlation (r=-.264, sig.=.462, p>.05) is not significant, the interpretation line in the figure shows that there appears to be a tendency that the earlier the test taker finished, the lower the score they obtained if more than 250 seconds (i.e. 42% of the total time allowance) were unused. This may suggest that test takers who used less than 60% of the time allocated underestimated the demand of the **Summarize Written Text** item.
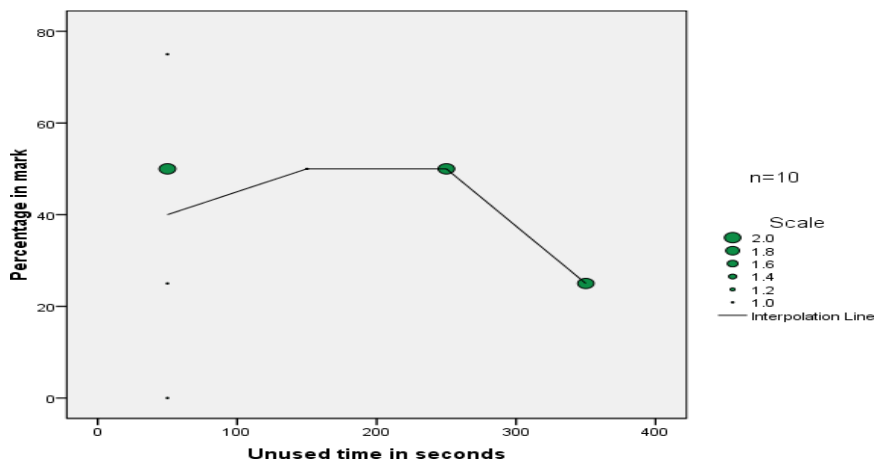

Figure 3: Correlation between unused time and the **Summarize Written Text** performance

5.2 Face validity of the reading-to-write and writing-only tasks

After completing the items and participating in the stimulated recall, all test takers were interviewed about their perceptions of the two item types.

5.2.1 Tasks and topics

Regarding the **Summarize Written Text** item, all test takers commented that they were not familiar with the topic, i.e. democracy. They found the source text difficult because it contained a lot of details, however most test takers said they could manage to understand the main ideas after reading it a few times because the level of the language was appropriate to them. They thought the item was not easy because they 'had to put the main idea of the whole paragraph into one sentence' (P7). It is worth noting that although all writers commented on the unfamiliarity of the topic, they seemed to be able to develop clear internal goals at an early stage and applied strategies to achieve them while composing the summary.

On the other hand, the test takers preferred the **Write Essay** item mostly because they were allowed to express their opinions on the topic, i.e. advertising. Regarding the topic, some said they were not familiar with it, while others expressed great interest in it. In the writing or testing literature, topic effect is one of the well-defined factors affecting writers' performance (Clapham, 1996). It seems that topic effect may affect test takers' composing process less on the **Summarize Written Text** than on the **Write Essay** item. It is possible that, on the **Summarize Written Text** item, test takers' interaction with the source text provided them with some knowledge of the topic. The topic effect on their writing process was thus decreased. However, topic effect on test takers' actual performance seems to be a more complicated issue. Table 4 shows that, on the **Summarize Written Text** item, all test takers claimed their unfamiliarity with the topic but their actual scores varied. On the other hand, while topic effect seems to hinder writers' composing process on the **Write Essay** item, such effect may not be clearly reflected in their actual scores. This is an interesting issue worthy of further investigation.

Demonstrating cognitive validity and face validity of PTE Academic writing items **Summarize Written Text** and **Write Essay**

11

Table 5: Familiarity with the topic and actual performance				
Participants	Summarize Written Text		Write Essay	
	Familiarity with the topic	Actual scores in %	Familiarity with the topic	Actual scores in %
1	Unfamiliar	50	Unfamiliar	70
2	Unfamiliar	75	Interesting	50
3	Unfamiliar	25	OK	100[5]
4	Unfamiliar	50	Unfamiliar	80
5	Unfamiliar	50	OK	70
6	Unfamiliar	50	Unfamiliar	40
7	Unfamiliar	0	Interesting	80
8	Unfamiliar	25	OK	70
9	Unfamiliar	25	Interesting	80
10	Unfamiliar	50	Unfamiliar	70

5.2.2   The connection with the real academic writing context

A majority of the test takers thought that the **Summarize Written Text** item simulated a more authentic academic writing context than the **Write Essay** item did. They believed the skills tested were similar to those they were required to perform in their studies: 'It's like doing something from different books that you have read. It's quite similar to what I experienced in academic situations' (P7). Two test takers (P2 and P8) thought that the two item types reflected different stages of their academic writing experiences. The **Summarize Written Text** item represents an initial stage where they have to understand and select ideas for the writing and summarize other people's work. The **Write Essay** item is similar to contexts in which they are expected to present their opinions on an issue in an organized way.

5.2.3   As a teaching item and as a testing item

More than half of the test takers thought that the **Summarize Written Text** item would be a better academic writing teaching task; firstly because they 'do not just write about agree[ing] or disagree[ing] with a topic' (P10) on their courses. Secondly, they thought summarizing skills were difficult to master, so that they wanted to learn the skills. A few writers preferred to do both items because they believed the items focused on different essential academic skills and they saw the connection between both item types.

When they were asked which item they preferred as a test item of their academic writing ability, test takers responded differently. Almost all participants were more confident with the **Write Essay** item based on the following reasons: Firstly, the **Write Essay** item was more interesting because it gave them more freedom to express themselves. This might motivate test takers to complete the item. However, the fact that test takers like writing about their ideas needs to be handled with caution. According to Weir (1993), the construct of writing can be differentiated into Interactional Operations and Informational Operations. While interactional operations represent writing purposes achieved in social and service texts, informational operations refer to those purposes usually achieved in academic texts. Expressing opinions, however, falls into the first category.

Secondly, the test takers believed that the **Write Essay** item could better assess their ability because 'writers had nothing to copy' (P10). The incidence of verbatim citations from the source text is not only the concern of the test takers but also of many researchers (e.g. Cumming et al., 2005; Lewkowicz, 1994 and Plakans, 2008). From the testing perspective, this is an issue only if the verbatim source text citations are not identified in the scoring procedure and not reflected in the scores.

Thirdly, the test takers believed not being able to understand the source text on the **Summarize Written Text** item caused problems. As mentioned earlier, according to the test takers, the linguistic level of the source text was appropriate, but the topic was unfamiliar to them. Most test takers mentioned that they had difficulty in understanding the text at the first attempt but after a few attempts they were able to get the general ideas. It is essential to ask whether poor performances on the summary item was due to deficiency in reading comprehension ability, discourse synthesis skills such as selecting relevant content from source texts and organizing contextual structure, and/or academic writing skills such as monitoring and editing.

[5] Due to a technical error, participant 3 exceeded the time allowance for writing-only item. Her score was not included in any statistical analysis carried out during this study.

Finally, the test takers commented that the ***Write Essay*** item allowed them to demonstrate their organizing skills in terms of how to arrange different paragraphs in a logical way but this was not tested on the Summarized Written Text item because they were required to summarize the source text in one sentence. However, while test takers can structure the summary by merely following the global organization of the source text, good summary writers always generate a superordinate category and organize content units to be subsumed into it (see Brown & Day, 1983).

5.3 Performances on the reading-to-write and writing-only items

This sub-section reports and discusses results regarding the test takers' performance on the ***Summarize Written Text*** and ***Write Essay*** items.

5.3.1 Actual performances

As marked by Pearson automated scoring technology, the test takers' mean score was 40%[6] on the ***Summarize Written Text*** item and 68% on the ***Write Essay*** item. The Wilcoxon test in the repeated measures was conducted and the result showed that their ***Write Essay*** scores were significantly (z=1.96 sig=.05, p<.05) better than their ***Summarize Written Text*** scores. The possible reasons for the poorer performance on the ***Summarize Written Text*** item are discussed below.

5.3.2 Poor performance on the *Summarize Written Text* item - low word requirement but high cognitive demand

The poor performance on the summary item was perhaps a result of the test takers' underestimation of the cognitive demand of the task. All participants in this study had intermediate writing ability (i.e. 6-7 IELTS writing scores) and the word requirement of the reading-to-task was intuitively unchallenging (i.e. one sentence). As mentioned earlier, most test takers submitted their text early on the ***Summarize Written Text*** item. Test takers believed that they 'had finished everything' (P8). A comparison between participants' perceived scores and the actual scores may also explain the issue. Table 6 below shows that nine out of ten test takers overestimated their scores on the summary performance while only three participants overestimated their ***Write Essay*** scores. Bosher (1998) argues that writers who performed badly on the reading-to-write task may not be well aware of the need to interact with the source text. But this does not seem to be the case in the present study because all test takers employed some discourse synthesis strategies while composing the summary item. Therefore, it appears that their poor results may be due to ineffective use of those strategies. Future studies should explore this issue further.

Participants	Summarize Written Text		Write Essay	
	Perceived scores in %	Actual scores in %	Perceived scores in %	Actual scores in %
1	60	50	40	70
2	60	75	70	50
3	60	25	70 or 80	100[7]
4	60	50	80	80
5	60	50	70	70
6	40	50	50	40
7	70	0	100	80
8	60 or 70	25	60 or 70	70
9	70	25	80	80
10	60	50	30	70

Table 6: Test takers' perceived performance and actual performance

[6] As the maximum item score a test taker can receive for each item type differs both item scores were standardized into percentage as the basis for comparison.

[7] See footnote 5

Demonstrating cognitive validity and face validity of PTE Academic writing items *Summarize Written Text* and *Write Essay*

13

5.3.3   Correlation between the reading-to-write and writing-only performances

Another issue this study aims to explore is the correlation between scores on the *Summarize Written Text* and *Write Essay* performances. Due to the small number of participants in this part, the correlation analysis was performed using additional scores from the Pearson data bank. Pearson provided the researcher with 200 scores[8] from 100 other test takers. After selecting test takers who had a similar profile as the participants in the present study, i.e. postgraduates and ESL learners,

50 scores form 25 test takers were used. The correlation was tested by the Correlation Coefficients Pearson test and the result shows that the correlation between the *Summarize Written Text* and *Write Essay* performances was not significant (r=.130; sig.=.462, p>.05). The result is parallel to Esmaeili's (2002) and Yu's (2008) findings. As discussed earlier, test takers in the present study apparently employed writing processes differently on the *Summarize Written Text* and *Write Essay* items. The cognitive and context constructs underlying both items are seemingly different. This is perhaps not an issue when the two items serve as sub-tasks which contribute to a final writing score. According to Alderson et al. (1995) correlations between the sub-tasks are expected to be low or non-significant when they are measuring different aspects of ability, i.e. academic writing ability in this case, but correlations between the sub-tasks and the whole test are expected to be high. Correlations between the *Summarize Written Text* and *Write Essay* scores and the overall writing scores in PTE Academic cannot be commented on here because six item types (15 items) in total contribute to the overall writing score.

6. Conclusion

This study demonstrates that the cognitive processes required to complete the *Summarize Written Text* and *Write Essay* items are not the same. Ultimately, most test takers employed a knowledge-telling approach to complete the *Write Essay* item. Their composing process seemed to rely heavily on retrieving relevant ideas from personal knowledge of the topic and metacognitive knowledge of the genre. In most cases, they composed the item in a rather linear manner by translating the retrieved content without much effort devoted to further organizing or elaborating them. Most test takers did not perform much macro-planning before writing. However, the item engaged test takers in considerable micro-planning while writing, monitoring and revising at low levels. The *Summarize Written Text* item, on the other hand, engaged test takers in a comparatively more recursive composing process. Most test takers engaged in careful macro-planning as well as some discourse synthesis processes such as organizing textual ideas in the source text and selecting relevant ideas from the source text. However, as the item requires test takers to write one sentence only, micro-planning, monitoring and revising do not seem to be activated much by this item.

The study provides cognitive validity evidence that different varieties of cognitive processes are elicited by the *Summarize Written Text* and *Write Essay* items. Nevertheless, problem-solving strategies involved in the knowledge transforming approach as well as high-level monitoring and revising which are usually employed by skilled writers do not seem to be sufficiently activated by the items. Previous research shows that comparison items tend to elicit monitoring and revising at an advanced level (e.g., Severinson Eklundh and Kollberg, 2003). It might be desirable to explore the possibility of either incorporating a comparison element into any of the existing item types or adding a new reading-to-write comparison item type to the PTE Academic Writing Section to allow test takers, especially those skilled ones, to employ such processes.

The findings of this study will hopefully provide insights relative to the cognitive validity of two writing item types. As test takers in this study performed significantly better on the *Write Essay* than the *Summarize Written Text* item. It might be useful to mention the cognitive processes required to complete both item types either on the score report or in the test manual. Nevertheless, in view of the limited data set used in this study, more research is necessary to confirm the preliminary findings.

In addition, this study has demonstrated that the use of screen video and stimulated recall can prove useful to researchers who are interested in investigating cognitive processes engendered by test items. The investigation of the composing process, especially on reading-to-write items, would be more complete if the eye-movement tracking technique could be incorporated. Keystroke logging would also provide richer and more accurate composing records and allow a more time-effective data analysis procedure.

---

[8] The additional 200 scores (100 per task) were taken from performances on the same *Summarize Written Text* and *Write Essay* items used in this study.

Demonstrating cognitive validity and face validity of PTE Academic writing items *Summarize Written Text* and *Write Essay*

References

Alderson, C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Asencion-Delaney, Y. (2008). Investigating the reading-to-write construct. *Journal of English for Academic Purposes*, *7*, 140-150.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillside, NJ: Lawrence Erbaum Associates.

Bosher, S. (1998). The composing processes of three Southeast Asian Writers at the post-secondary level: an exploratory study. *Journal of Second Language Writing*, *7*(2), 205-241.

Brown, A.L., & Day, J. D. (1983). Macrorules or summarizing texts: The development of expertise. *Journal of Verbal Learning and Verbal Behavior*, *22*, 1-14.

Carson, J. (2001). A task analysis of reading and writing in academic contexts. In D. Belcher & A. Hirvela (eds)., *Linking literacies: Perspectives on L2 reading-writing connections* (pp. 246-270). Ann Arbor, MI: University of Michigan Press.

Campbell, C. (1990). Writing with others' words: using background reading text in academic compositions. In B. Kroll (Ed.), *Second language writing research insights for the classroom* (pp. 211-230). New York: Cambridge University Press.

Clapham, C. (1996) *The development of IELTS: A study in the effect of background knowledge on reading comprehension*. Studies in Language Testing 4. Cambridge: UCLES/Cambridge University Press.

Cumming, A., Kantor, R., Baba, K., Eedosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing, 10*, 5-43.

Cumming, A., Rebuffot, J., & Ledwell, M. (1989). Reading and summarizing texts in first and second languages. *Reading and Writing, 2*, 201-219.

Esmaeili, H. (2002). Reading-to-write reading and writing tasks and ESL students' reading and writing performance in an English language test. *Canadian Modern Language Review, 58*, 599-622.

Grabe, W., & Kaplan, F. L. (1996). *Theory and Practice of Writing: An Applied Linguistic Perspective*. London: Longman.

Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 2000 - Writing: Composition, community and assessment.* TOEFL Monograph Report. Princeton, NJ: Educational Testing Service.

Hayes, J., & Flower, L. (1983). Uncovering cognitive process in writing: An introduction to protocol analysis. In P. Mosenthal, L. Tamor, & S. Walmsley (Eds.), *Research on writing: Principles and methods* (pp. 206-219). New York: Longman Inc.

Horowitz, D. M. (1986). What professors actually require: Academic tasks for the ESL classroom. *TESOL Quarterly, 20*, 445-460.

Kowal, S., & O'Connell, D. (1987). Writing as language behaviour: Myths, models, methods. In A. Matsuhashi (Ed.), *Writing in real time* (pp. 108-132). Norwood, NJ: Ablex Publishing Corporation.

Lewkowicz, J. (1994). Writing from sources: Does source material help or hinder students' performance? In M. Bird et al (eds.), *Language and Learning* (pp. 204-217). Hong Kong: Government Printer.

Matsuhashi, A. (1982). Explorations in the real-time production of written discourse. In M. Nystrand (ed.) *What writers know* (pp. 269-290). New York: Academic Press.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Pearson. (2009). *The Official guide to Pearson Test of English Academic*. UK: Longman.

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

Read, J. (1990). Providing relevant content in an EAP writing test. *English for Specific Purposes, 9*(2), 109-121.

Demonstrating cognitive validity and face validity of PTE Academic writing items *Summarize Written Text* and *Write Essay*

15

Severinson Eklundh, K., & Kollberg, P. (2003). Emerging discourse structure: computer-assisted episode analysis as a window to global revision in university students' writing. ***Journal of Pragmatics, 35***, 869-891.

Shaw, S. and Weir, C. J. (2007). ***Examining writing: Research and practice in assessing second language writing***. Studies in Language Testing. Cambridge: Cambridge University Press.

Spivey, N. N. (1984). ***Discourse synthesis: Constructing texts in reading and writing***. Outstanding Dissertation Monograph. Newark, DE: International Reading Association.

Weigle, S. C. (2002***). Assessing writing***. Cambridge: Cambridge University Press.

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English, ***Assessing Writing***, 9(1), 27-55.

Weir, C. J. (1993). ***Understanding and developing language tests***. New York: Prentice Hall.

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

Yu, G. (2008). Reading to summarize in English and Chinese: A tale of two languages? ***Language Testing***, 25(4), 521-551.

16

Demonstrating cognitive validity and face validity of PTE Academic writing items *Summarize Written Text* and *Write Essay*