Validating the integrated writing task of the TOEFL internet-based test (iBT): Linguistic analysis of test takers' use of input material

Nao OHKUBO

Graduate School of Tsuda College, Japan

Abstract

Launched in 2005, the integrated writing task of the TOEFL iBT was designed to measure a test taker's readiness to use English in an academic context. The aim of the study reported in this paper was to provide further evidence to justify that the test score on the integrated writing task reflects the test taker's ability to produce academic writing.

This study analysed the performance of six test takers with respect to two aspects of language use: acknowledging and reformulating information from source texts. Different aspects of test taker performance were identified through an analysis of the language features in test takers' written texts and a post-task interview with the test takers. The results showed that test performance mostly conformed to the test construct. Two successful test takers (as measured by scores assigned to their writing performance) were able to display skills of attributing information to the input texts as required by the task, while three less successful test takers lacked these skills. This was interpreted as evidence that test performance elicited by the integrated writing task was closely linked to the skills required in academic contexts.

The analysis of test takers' performance with respect to reformulating source texts, however, showed that the task did not elicit all the language skills required for successful performance in the academic domain. These instances of construct underrepresentation could thus be regarded as threats to task validity. Given the small sample size used for the current research, this paper advocates that further research be conducted in the interest of establishing the validity of the integrated writing task as a measure of academic language ability.

Melbourne Papers in Language Testing 2009 Volume 14.1 pp. 1-31. The Language Testing Research Centre, The University of Melbourne.

1. Introduction

The TOEFL (Test of English as a Foreign Language) is an English language test used for the admission of non-native English speakers into tertiary institutions in North America and other English-speaking countries. Previously, the TOEFL writing component contained only an independent task. However, this task was criticised because it did not closely resemble the genres used in real academic settings (Hale et al., 1996). To remedy the discrepancy between the test and the target language use situation, the integrated writing task was launched in the writing component of the TOEFL iBT in 2005.

Test designers initially developed two prototype integrated writing tasks, which used a single source text (a reading passage or a lecture), but this plan was abandoned for two reasons (Enright et al., 2008; Pearlman, 2008). First, test takers relied heavily on lifted expressions from the reading passage without quotation marks or acknowledgement of authors (Cumming et al., 2006; Enright et al., 2008; Lumley & Brown, 2006). Second, it would take too long for the test takers to perform two types of the integrated writing tasks (reading then writing, and listening then writing) (Pearlman, 2008). Therefore, two source texts are used for the current form of the integrated writing task instead of one source text.

However, this integrated writing task has rarely been the subject of validation studies by the Educational Testing Service (ETS) (Carrell, 2007; Cumming et al., 2006; Cumming, Kantor, Powers, Santos, & Taylor, 2000; Lee & Kantor, 2005; Lumley & Brown, 2006; Sawaki, Stricker, & Oranje, 2008). Therefore, it is not clear on what basis the distinction is made between successful and less successful test takers. The current study seeks to provide further evidence of the validity of the integrated writing task by investigating how closely test takers' language use matches the integrated writing task construct.

2. Approaches to Validating Writing Tasks

Construct validation is concerned with whether inferences from performance in the test situation to performance in the target domain can be justified (Messick, 1989). Kane's (2001) interpretive argument approach has been applied as the validation framework for the TOEFL iBT (See for details, Chapelle, Enright, & Jamieson, 2008; Kane, 2001; Kane, Crooks, & Cohen, 1999). Under the current TOEFL validation framework, the link between test performance and target domain performance can be connected by a chain of inference via the test construct. The justification for this chain of inference should be logical and supported by defensible sources of evidence (Kane, 2006; Kane et al., 1999; McNamara & Roever, 2006; Messick, 1989).

One of the approaches used to justify these inferences is to investigate how well linguistic characteristics elicited from test performance accord with the test construct (Cumming, 2008; Cumming et al., 2006; Huff et al., 2008; Weir, 2005; Xi, 2008). This discourse analytic approach was used in Lumley and Brown's (2006) study. Lumley and Brown investigated how 60 English as a second language (ESL) and English as foreign language (EFL) students reformulated source texts by using two prototype (reading then writing) integrated writing tasks. They found that the high-scoring test takers could use their own words while the less successful test takers only modified the occasional word and otherwise left original sentences unchanged.

However, while Lumley and Brown (2006) used a prototype (reading then writing) integrated writing task, the current TOEFL iBT integrated writing task requires test takers to use information from both reading and listening texts. It is uncertain how the different modes of the input texts will affect the ways test takers cope with writing (Cumming et al., 2006).

When completing the TOEFL iBT integrated writing task, test takers are required to summarise information from a 200 - 300 word text and a two-minute, 150 - 250 word lecture. Though writing is judged from

two perspectives, content and language use (Cumming, 2001; Cumming et al., 2000), the current study investigates language use only. In particular, it will focus on two unique linguistic features: 1) acknowledging and 2) reformulating information in the source texts (Cumming et al., 2006; Cumming et al., 2000).

One key language requirement of academic writing is the ability to acknowledge the input texts as the source of information. It is worth noting that students' ability to explicitly attribute information in the written texts to input sources has been identified in the EAP literature as an important component of academic literacy (Carkin, 2005; Hinkel, 2002; Hyland, 2006; Oshima & Hogue, 2005; Swales & Feak, 2004). However, ESL/EFL students usually struggle to acknowledge the source texts because of their cultural, language, and educational background or insufficient English proficiency (Barks & Watts, 2001; Bloch, 2001; Cumming et al., 2006; Currie, 1998; C. Thompson, 2006). Though the need to identify input sources is not explicitly mentioned in the *TOEFL Tips* or *the Official Guide to the New TOEFL iBT*, the fact that this skill is modelled in the higher level benchmark essays can serve as evidence that the construct of the TOEFL iBT writing paper reflects the skills required in the target language use situation.

Another part of the integrated writing task construct is the ability to reformulate the source texts. In academic situations, if test takers use expressions that are similar to those used in the original text without quotation marks, their response is regarded as plagiarism (Barks & Watts, 2001; Bloch, 2001; Currie, 1998; Pecorari, 2001; C. Thompson, 2006). On the other hand, when test takers restate the ideas from the source materials by using their own words, their response is characterised as paraphrasing. Paraphrasing is strongly recommended for academic writing because direct citation does not reveal whether students actually understand the source texts (Hirvela, 2004; Keck, 2006; Swales & Feak, 2004). Yet, ESL/EFL students often replace two or three words from an original sentence and copy the rest of the sentence when they attempt to reformulate source texts (Campbell, 1990; Hirvela, 2004; Keck, 2006; Shi, 2004). In other words, their paraphrasing is too similar to the original text.

The question of whether such minimal paraphrasing is acceptable in the target academic domain is not a simple one, however. While Swales and Feak (2004) stated that students should use their own words in order to avoid plagiarism, some scholars argue that the use of language that closely approximates that of the original text might be acceptable given that ESL students are in a period of transition from novice to expert academic writers (Canagarajah, 2002; Howard, 1995, 1999; Pennycook, 1996). Furthermore, C. Thompson (2006) showed that individual lecturers differ in their attitudes toward copying. Sutherland-Smith (2005) indicated that the charge of plagiarism is not actually filed in many cases due to the fact that detecting plagiarism is such a time-consuming process. Thus, thorny issues surround the question of what kind of relationship between input texts and students' academic writing is acceptable.

Despite these uncertainties, the documentation relating to the TOEFL iBT suggests that the ability to reformulate the information from the source materials successfully in one's own words is deemed important for successful performance on the integrated writing task (Cumming et al., 2000; Educational Testing Service, 2007a, 2007b). Both score rubrics and the *TOEFL Tips* warn that "test takers receive a score of zero if all they do is copy words from the reading passage" (Educational Testing Service, 2007b, p.33).

Based on what is reported above, two aspects of the construct of the integrated writing task can be articulated:

1. The test taker's ability to explicitly attribute information in their essay to the input texts and;

2. The test taker's ability to reformulate information in the input texts in their own words to avoid plagiarism.

3. Research Questions

The main purpose of the current study is to validate the integrated writing task of the TOEFL iBT with respect to language use via a

discourse analysis of successful and less successful test takers' performance. A research question with two aspects was addressed:

> How different are successful and less successful test takers' responses to the integrated writing task in terms of:
>
> 1) appropriate attribution of information drawn from the input texts and;
>
> 2) use of students' own words to reformulate information in the input texts?

4. Methodology

Since a scant number of studies have used the current form of the integrated writing task, this study consists of an in-depth analysis of a small number of test-takers' performances by using a limited number of participants.

For qualitative research, triangulation of data collection methods and detailed description of the data is necessary to make the data interpretation credible, transferable, and dependable (Creswell, 2007; Denzin & Lincoln, 2000; Lichtman, 2006; Mackey & Gass, 2005; Miles & Huberman, 1994). Therefore, this study used not only test takers' written texts but also interviews with the test takers. The quality of test performance identified by these methods was then matched to the test construct.

4.2. Participants

4.2.1. Test Taker Participants

Six students participated in the study. Their backgrounds are shown in Table 1. The participants' real names are replaced with pseudonyms for ethical reasons.

Name	Gender	Level	Field of Study	First language
Miri	F	PhD	Applied Linguistics	Korean
Ben	M	MA	Physiotherapy	Japanese
Hana	F	Study Abroad[1]	Politics (Faculty of Arts)	Japanese
Ali	M	Language school	IT	Arabic
Xavier	M	Language school	Marketing	Portuguese
Lina	F	PhD	Science Education	English/ Chinese

Table 1. Background of Test Participants

4.2.2. Rater Participants

Three experienced raters were recruited to rate the integrated writing tasks. Their background is shown in Table 2.

	Rater 1	Rater 2	Rater 3
Gender	F	F	F
Teaching experience (years)	13	25	5

Table 2. Backgrounds of Raters

[1] The study abroad program consists of six-month academic English courses at a language school and one-term undergraduate-level study.

	Rater 1	**Rater 2**	**Rater 3**
Highest degree	PhD	MA	PhD
Rating experience (years)	DELA[2] - 3	IELTS – 5 DELA – 10 OET[3]	IELTS – 7 DELA - 6

Table 2. Backgrounds of Raters (continued)

4.3. Instruments

4.3.1. Task

The current study used a practice test for the integrated writing task from *the Official Guide to the New TOEFL iBT* (Educational Testing Service, 2007a). This task requires test takers to "summarise the points made in the lecture you just heard, being sure to specifically explain how they cast doubt on points made in the reading" (Educational Testing Service, 2007a, p. 284). While this practice task was not an official one prepared by the ETS, the task was designed to replicate an operational version as closely as possible.

In the integrated writing task, the reading, listening and writing components are centred on a single academic topic. The reading text usually consists of one main idea and three supporting ideas. Comments that either support or oppose these ideas are provided by the lecture.

[2] DELA is the Diagnostic English Language Assessment. This test assesses English proficiency of non-native English speakers starting study in Australia.

[3] OET is the Occupational English Test. The OET assesses the English proficiency of overseas-qualified health professional workers who plan to work in Australia.

4.3.2. The Integrated Writing Rubrics

A score between 0 and 5 is given by the raters according to the Integrated Writing Rubrics (Appendix A) (Educational Testing Service, 2005). Though the test score is reported as a holistic score, a test taker's response is assessed from two perspectives: content and language use (Cumming et al., 2000). The perspective of content embodies criteria for judging a test taker's ability to select important information from the reading text and the lecture, as well connecting information from the lecture to relevant information from the reading text. When some information from both the reading text and the lecture is included in the writing task, at least a score of 2 will be given (Pearlman, 2008). If all the necessary information from these two input materials is included, a score of 4 or 5 will be given (Educational Testing Service, 2007a). In terms of language use, organisation, grammatical structures and expressions are also assessed. The rubric shows that a score of 4 or higher can be given when the test taker's response does not include frequent and noticeable errors in his or her writing.

4.3.3. Benchmark Essays

The benchmark essays provide distinctive characteristics of test takers' performance associated with each score level. Five benchmark essays are provided, according to each score from 1 to 5, in *the Official Guide to the new TOEFL iBT*. These benchmark essays attach samples of raters' comments on the test takers' performance. These comments show brief reasons for raters' test scores.

4.3.4. The Interview Questions

This study employed interviews because they offer a means to elicit test takers' reflection on specific performances during the test such as reformulation of the source texts. Semi-structured interviews were used here because the researcher could ask test takers consistent questions, as well as adding or deleting questions depending on test-takers' response (Mackey & Gass, 2005). The questions were first asked in a closed format (e.g. Do you feel that you have copied any

chunks from the listening input without changing them?), and then in an open-ended form (e.g. Why or why not?). The semi-structured interviews consisted of 12 questions on time limitation, reformulation of the source texts, etc. An example of a question on time limitation was "You had 20 minutes for your response. Was there enough time?" Questions on reformulation of the source texts included "Have you added any of your own ideas?", "Did you find it hard to use your own words?" and "So you used your own words. Did you paraphrase?"

4.4. Data Collection Procedures

4.4.1. Test Taker Participants

On the day of the data collection, the test takers participated in a brief tutoring session on the integrated writing task and completed the writing task. The pilot study showed that test participants required tutoring in order to familiarise themselves with the tasks. All instructions and the interview were conducted in English.

In order to complete the integrated writing task, test taker participants first read a text that was based on an academic topic (230-300 words) for three minutes. The test takers then listened to a two-minute lecture related to the reading text (230-300 words). After this, they wrote a summary for twenty minutes based on the text and lecture. The test takers were allowed to take notes while reading and listening, and to look at the reading text while writing.

After completing the integrated writing task, the test takers were asked to participate in a semi-structured interview about their performance on the integrated writing task.

4.4.2. Rater Participants

Rater training was conducted to instruct the raters on how to assess the integrated writing task according to *the Integrated Writing Task Rubrics* and benchmark essays. Because of the raters' busy schedules, each rater scored the six texts produced by the test takers in their own

time and returned the scores to the researcher within a week of the rater's training.

4.5. Data Analysis Procedures

4.5.1. Test score

The test scores assigned by the raters are shown in Table 3. The inter-rater reliability among the three raters, as assessed using Cronbach's alpha, was 0.98. This is considered to be highly reliable for rating written texts (Weigle, 2002). An average of the three raters' scores was used to arrive at the final score.

Based on the final score, the test takers were classified as successful or unsuccessful test takers to investigate whether the successful test takers were more able to display skills as required by the task, while the less successful test takers lacked these skills. A writing score of three was deemed to be a borderline score for entering universities (e.g. The University of Melbourne, 2007a, 2007b). Those scoring over a three (Lina and Hana) were therefore regarded as successful test takers and those scoring below three (Xavier, Miri and Ali) were regarded as unsuccessful for the purposes of the current study. As for Ben, his score cannot be regarded as either successful or unsuccessful score. Therefore, his response is seen as a borderline case and will not be analysed in depth.

Test takers	Rater 1	Rater 2	Rater 3	The final score
Lina	5	5	5	5
Hana	4	4	4	4
Ben	3	3	3	3
Xavier	2	2	3	2.3
Miri	2	1	2	1.7
Ali	1	1	2	1.3

Table 3. The scores given by three raters

4.5.2. Procedures for the Discourse Analysis

The discourse analysis deals with the use of discourse devices which acknowledge the input texts as the source of the information in test takers' writing and reformulate the information in the source texts. The first analysis was designed to identify the discourse devices used to acknowledge the input source. The ability to use these devices is analysed according to the system applied by Cumming et al. (2006) and Cumming et al. (2000). In their system, the reporting verbs are used for identifying "who or what is presented as the source of the language being reported" (Cumming et al., 2006, p. 65). If the reporting verbs are the same or similar reporting verbs as shown in G. Thompson and Yiyun (1991), these verbs were selected for analysis in the current study.

The second analysis focused on reformulations of the information in the source texts. A preliminary analysis showed that paraphrase analysis used by Lumley and Brown (2006), or counting the number of verbatim sentences (see Cumming et al., 2006; Keck, 2006) was not suitable for this study. Although the participants in those studies extensively copied the original texts, none of the test takers in the current study copied more than three consecutive words. A plausible explanation for this is the different nature of the integrated writing tasks in each case. As suggested in Cumming et al. (2006), different input modes can affect the amount of copying from the input texts. Therefore, a different approach from that used by Lumley and Brown (2006) was necessary to analyse reformulation of the information in the source texts in the current study. This involved identifying the linguistic characteristics of the source texts, and comparing these characteristics in the produced texts with those of the benchmark essays.

4.5.3. Procedures for analysing post-task interview

Interviews with the test takers were conducted to collect test takers' reflection on reformulation of the information in the source texts. Test-takers' comments on the questions were recorded and

transcribed by the researcher. Key points of these comments were used for the current analysis.

5. Findings

5.1. The Test Takers' Written Text

The first linguistic analysis of the test takers' written texts deals with the use of discourse devices which acknowledge the input texts as the source of the information in test takers' writing. An analysis of the benchmark essays (Table 4) reveals that the high-scoring test takers (with a score of four or five) use various discourse devices to acknowledge the input texts as the source of their information. For example, the high-scoring test takers could acknowledge an author ("the lecture", "Professor") and use appropriate reporting verbs ("refutes", "pointed out"). On the other hand, the less successful test takers (with a score of two or one) made scant use of these kind of devices.

Score	Identification of the source evidence
5	"The lecture completely refutes", "It is said in the lecture", "Contrary the belief in the passage", "the professor says", "The lecture refutes", "Professor also offers", "She says"
4	"The lecture warned", "The lecture also pointed out that"
3	"The lecture might make the reader doubt"
2	NA
1	"This lecture said"

Table 4. Identification of the source evidence in the benchmark essays

Table 5 shows the discourse devices used by the test takers in the current study to attribute the information in their essays to the reading or listening input texts. As shown in the analysis of benchmark essays in Table 4, the high-scoring test takers (Lina, Hana)

could also acknowledge the source texts ("The lecture", "the lecturer") in their essays. Furthermore, they could use a variety of reporting verbs ("doubts on", "suggest", "respond"). On the other hand, the less successful test takers (Miri, Ali) could not identify the source evidence in their essays.

Test takers	Score	Identification of the source evidence
Lina	5	"The lecture doubts on", "unlike what is stated in the text", "The lecturer seems to suggest", "contradicting the argument in the text that suggest", "the lecturer suggests"
Hana	4	"The lecture explained", "the lecture doubted", "it argued", "the lecture responded"
Miri	1.7	NA
Ali	1.3	NA

Table 5. Identification of the source evidence in this study

It can be seen from this analysis that, just as in the benchmark essays, the successful test takers used a range of devices to identify the source of their information while the less successful test takers failed to use any of these discourse devices. This finding indicates that successful test takers explicitly attribute information in their essays to the input texts, while less successful test takers do not.

The second linguistic feature investigated in this study is reformulation of the information in the source texts and compared those to the benchmark scripts. Reformulated expressions drawn from part of the lecture were compared between the high-scoring and low-scoring essays as shown in Table 6. The misspellings and grammatically incorrect language the test takers used in their texts have not been corrected here in order to show what the test takers actually wrote.

The successful test takers' essays	
Synonyms	
Input text	gain an increased sense of self-worth
Lina	"to raise the self-worth"
Input text	receive appreciation
Benchmark essay 5	"gain appreciation"
Syntactic changes	
Input text	receive appreciation
Hana	"could be appreciated by someone"
Input text	gain an increased sense of self-worth
Benchmark essay 4	"his/her self-worth increase"
Similar grammatical structure to the original texts	
Input text	the donor receive appreciation and approval from the stranger and society
Lina	"receives approval from the recipient and society at large"
Input text	a person donates a kidney to a relative, or even to a complete stranger
Benchmark essay 4	"a man give one of his/her kidney to a family member or even a stranger"

Table 6. Examples of paraphrases elicited from the texts by successful test takers and in the high-scoring benchmark essays

Neither the successful test takers' essays nor the high scoring benchmark essays (4, 5) showed any evidence of direct copying from the lecture. Instead, they used synonyms or minor syntactic changes

(e.g., from verb to noun or from noun to verb), while preserving grammatical structures which were the same as or similar to the original texts. This comparison showed that while the successful test takers did not use the same words as those used in the input texts, their texts were nevertheless closely based on these sources.

Interestingly, the less successful test takers (Xavier and Ali) and the writer of benchmark essay 2 were more likely to use their own words rather than make efficient use of the words in the input texts. In particular, Xavier and the writer of benchmark essay 2 used completely different words to define the meaning of "receive appreciation and approval" (Table 7).

The less successful test takers' essay	
Input texts	A selfless act, right? But…doesn't the donor receive appreciation and approval from the stranger and from society? Doesn't the donor gain an increased sense of self-worth?
Xavier	"because he want to look like a good person, or want respect from the others"
Ali	"this could give the donor a kind of self respect and he could gain some kind of appreciation without the natural greed within people"
Benchmark 2	"person expects the family of the person that has received the organ to give him or her thanks because of that favour"

Table 7. Examples of paraphrases elicited from the texts by less successful test takers and the low-scoring benchmark essays

Thus, the successful test takers were able to find synonyms and utilise syntactic changes, but they were not always able to restate the input texts in their own words. In contrast, the less successful test takers were more likely to use their own words.

5.2. Test Takers' Comments during the Interviews

The key points made in the post-test interviews about reformulation of the information in the input texts are shown in Table 8. The key points are classified according to reformulation of the information in the input texts or time limitation. Unfortunately, the interview with Lina was not recorded due to technical problems.

Test-takers (Test score)	**Key points of the interviews**
Reformulation of the information in the input texts	
Hana (4)	- I only copied a few words: altruism and valuable. - I paraphrased the words by using synonyms (e. g. from donate to donation), switching between active voice and passive voice and exchanging from subject to object.
Ben (3)	- I copied the key word altruism but I tried to paraphrase from the reading passage. - In the lecture, I copied words such as non-material valuable things and donation, but I could not remember the exact words in the lecture, so I just wrote some key words only.
Xavier (2.3)	- I copied a bit of the definition about altruism. - I used my own words. Main words must be the same but, in other ways, I tried to use my own words.
Miri (1.7)	- Copying from the original input was more difficult because I could not memorise all of the expressions in the input text. - I paid more attention to the content than to expressions used in the input texts. - I did not want to copy, so I tried to find synonyms or different words.

Table 8. Key findings from the interview

Test-takers (Test score)	**Key points of the interviews**

Reformulation of the information in the input texts	
Ali (1.3)	- I used my own words and paraphrased but I was not sure. - If the input was only reading, I clearly know whether I copied or paraphrased. -I don't know whether I copied from the lecture because I think that I tried to understand what the lecture was about. So I did not care about the words used in the lecture and I was just interested in the meaning.
Time limitation	
Hana (4)	- I think that I may be able to find more complicated grammatical structure or another structure if there is no time limitation.
Miri (1.7)	- Twenty minutes was a short time to think about how to paraphrase the original text in my own words.

Table 8. Key findings from the interview (continued)

This interview indicated that successful test takers, like Hana, were likely to focus on the language delivered in the input texts as well as the content. In spite of her attention to linguistic expressions in the input texts, she stated that she could not reformulate these expressions in her own words because of the limited time. In other words, minor modification is all that she could manage under the time constraints. On the other hand, the less successful test takers like Miri and Ali seemed to pay less attention to the language that was used in the input texts than to the content of the texts.

6. Discussion

This section addresses how closely the test-takers' performance identified in the current study reflects the test construct of the integrated writing task. It also explores any construct underrepresentative or irrelevant behaviours, which could be viewed as threats to task validity. Table 9 below summarises what this study found about the relationship between the construct of the integrated writing task as articulated in test documentation and in the relevant

research literature on academic writing and characteristics of the test takers' responses.

Construct of the integrated writing task	**Characteristics of the test takers' responses**
The ability to explicitly attribute information in their essay to input sources	It can be seen that the successful test takers were able to appropriately attribute the information in their essay to the relevant input texts.
The ability to reformulate information in the input texts in one's own words to avoid plagiarism	The successful test takers were able to pay closer attention to linguistic features in the input texts and to use synonyms and make some syntactical changes to the source text instead of copying from the source texts. However, they did not always use their own words. On the other hand, the less successful test takers were more likely to use their own words.

Table 9. The construct of the integrated writing task of the TOEFL iBT and test taker's response

With respect to the attribution of information drawn from the input texts, there was found to be a match between the performance of successful test takers and the integrated task construct. As previous studies have shown, explicitly acknowledging input sources plays an important role in writing research-based essays (Hinkel, 2002; Hyland, 2006; Swales & Feak, 2004). In other words, the successful test takers are likely to cope well with the writing demands in academic contexts. On the other hand, the less successful test takers, who were unable to make clear the boundaries between different texts, may face difficulties completing research-based essays and may need more preparation in this area before entering universities.

As far as reformulation of the information in the source texts is concerned, test taker performance and the integrated writing construct did not seem to be closely linked. While the successful test takers were better able to make minor grammatical changes of sentences used in the input texts, they were not always able to restate the input texts entirely in their own words. In fact, the less successful test takers were more likely to do this. In this sense, there was a mismatch between test performance and the test construct because EAP-related research has shown that high-proficiency students were able to avoid copying the input texts and to use their own words (Campbell, 1990; Currie, 1998; Fox, 2001; Hirvela, 2004; Keck, 2006; Shi, 2004). This mismatch will be discussed from three perspectives.

Firstly, most of the relevant EAP-related studies on reformulation of the information in the source texts have investigated the cases where reading texts were used as input. However, the introduction of listening may place additional demands on test takers (Cumming et al., 2006). The integrated writing construct, with respect to reformulation of the source texts, needs to be re-examined by EAP-related studies. For example, research should be done on how university students use information from lectures to prepare for writing an assignment. If the newly-articulated construct indicates that the successful students are likely to use synonyms instead of their own words in summarising the lectures, the behaviour elicited in response to the integrated writing task in the current study might be seen as valid.

Another possible explanation for the mismatch between test performance and the test construct is that the test situation places constraints on what is feasible to measure. The EAP-related studies pointed out that attitudes regarding the acceptability of plagiarism may be affected by the difficulty that teachers (or raters of essays) have in judging whether the students had copied or not (C. Thompson, 2006), as well as the time required for detecting plagiarism (Sutherland-Smith, 2005). However, under test conditions, such uncertainty is not acceptable for reasons of reliability and practicality, which are important aspects of validity for a large-scale and high-stakes examination, such as the TOEFL iBT (Bachman &

Palmer, 1996). Therefore, the integrated writing task was designed expressly to prevent such copying behaviour (Pearlman, 2008). In fact, the interviews with the test takers showed that they were unable to copy phrases from the lecture. The text analysis also showed that the test takers copied only key words in the input texts. This evidence indicates that the test takers' use of the lifted phrases was avoided in the current integrated task. However, copying is an unavoidable issue for academic writing in the target language use situation (Buranen, 1999; Currie, 1998; Howard, 1995; Pecorari, 2001; Pennycook, 1996). This suggests that an integrated task that prevents copying behaviour might be invalid in terms of authenticity, even if it is valid in terms of practicality and reliability.

The last issue related to this mismatch is how reformulation of the input texts exhibited by the test takers was evaluated by the raters of the current study. The raters assigned low scores to the test takers who were likely to use their own words, but who used incoherent and imprecise language. On the other hand, they gave high scores to the test takers who frequently used minor modifications, such as synonyms, but who wrote in appropriate and accurate language. Although such minor modifications seemed to be close to the original text, nevertheless, the latter test takers' texts were highly evaluated.

However, it is possible that academic lecturers might regard the texts written by the successful test takers of the current study as too close to the original texts (Bloch, 2001; Howard, 1995; Pecorari, 2001, 2003). If university students copy source texts, or if their writing is too similar to the source texts, their performance is regarded as plagiarism and they might receive warnings from their lecturers (Currie, 1998; C. Thompson, 2006). Thus, if it is true that paraphrasing is valued or accepted on the integrated writing task but it is not in fact predictive of test takers' ability to reformulate the information from the input texts in the target language use situation, this can be seen as a threat to the construct validity of the test.

These three issues indicate that further research is necessary to determine whether the integrated writing task is in fact a valid

representation of real world language use in respect to reformulation of the information in the source texts.

7. Limitations

This study was based on data from only six test takers and three raters. This small sample size means that the results reported here cannot be generalised. Future studies should therefore attempt to increase the number of participants.

Another limitation is task familiarity. As Lumley and Brown (2006) demonstrated, task familiarity affects test performance. The test takers in this study were not real test takers. Although a short orientation to the integrated writing task was provided to the participants, they were relatively unfamiliar with the task. Real test takers would be far more likely to have practised the integrated writing task.

Furthermore, although the task was designed to replicate an operational version as closely as possible, the task was not a real one. And while the raters used in this study were experienced and reliable, they were not official ETS raters. Therefore, the results of this study, based on an unofficial task and non-ETS raters, must be cautiously interpreted.

8. Conclusion

This study explored the validity of the new TOEFL iBT integrated writing task in terms of how closely the response of six test takers conformed to the test construct with respect to explicitly attributing the information in the test taker's texts to the source materials, and reformulating the information in the source texts. The results showed some instances of construct underrepresentative behaviour influenced by test conditions such as the time constraint. This construct underrepresentative behaviour might be seen as a threat to task validity because the task does not elicit all the skills required for successful performance in the academic settings such as restating the information in the input texts in students' own words. In order to

elicit these skills accurately under a test situation, it is recommended that future validity studies on the integrated writing task should re-examine the construct or collect more data on test takers' use of reformulation of the information in the input texts.

Acknowledgements

I would like to express my gratitude to Dr. Cathie Elder, the six test taker participants, and the three raters who generously volunteered their time and effort to take part in the study.

References

Bachman, L., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Barks, D., & Watts, P. (2001). Textual borrowing strategies for graduate-level ESL writers. In D. Belcher & A. Hirvela (Eds.), *Linking literacies: Perspectives on L2 reading-writing connections*. Ann Arbor, MI: The University of Michigan Press.

Bloch, J. (2001). Plagiarism and the ESL students: From printed to electronic texts. In D. Belcher & A. Hirvela (Eds.), *Linking literacies: Perspectives on L2 reading-writing connections*. Ann Arbor, MI: The University of Michigan Press.

Buranen, L. (1999). "But I wasn't cheating": Plagiarism and cross-cultural mythology In L. Buranen & A. M. Roy (Eds.), *Perspectives on plagiarism and intellectual property in a postmodern world* (pp. 63-74). NY: State University of New York.

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). Cambridge, UK: Cambridge University Press.

Canagarajah, S. (2002). Multilingual writers and the academic community: towards a critical relationship. *Journal of English for Academic Purposes, 1*, 29-44.

Carkin, S. (2005). English for Academic Purposes. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 85-98). Mahwah, NJ: Lawrence Erlbaum Associates.

Carrell, P. (2007). *Notetaking strategies and their relationship to performance on listening comprehension and communicative assessment tasks*. Princeton, NJ: Educational Testing Service.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In C. A. Chapelle, M. K. Enright & J. M. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 1-25). NY: Routledge.

Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches*. Thousand Oaks, CA: Sage Publications.

Cumming, A. (2001). ESL/EFL instructors' practices for writing assessment: specific purposes or general purposes? *Language Testing, 18*(2), 207-224.

Cumming, A. (2008). Assessing oral and literate abilities. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Language testing and assessment* (2nd ed.), Vol. 7, pp. 3-18. New York: Springer.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL*. Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper*. Princeton, NJ: Educational Testing Service.

Currie, P. (1998). Staying out of trouble: Apparent plagiarism and academic survival. *Journal of Second Language Writing, 7*(1), 1-18.

Denzin, N. K., & Lincoln, Y. S. (2000). Introduction: The discipline and practice of qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed.), pp. 1-28. Thousand Oaks, CA: Sage publications.

Educational Testing Service. (2005). *TOEFL iBT tips: How to prepare for the next generation TOEFL test and communicate with confidence*. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (2007a). *The official guide to the new TOEFL iBT*. New York: McGraw-Hill.

Educational Testing Service. (2007b). *TOEFL iBT Tips: How to prepare for the TOEFL iBT*. Princeton, NJ: Educational Testing Service.

Enright, M. K., Bridgeman, B., Eignor, D., Kantor, R. N., Mollaun, P., Nissan, S., et al. (2008). Prototyping new assessment tasks. In C. A. Chapelle, M. K. Enright & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 145-186). New York: Routledge.

Fox, J. D. (2001). *It's all about meaning: L2 test validation in and through the landscape of an evolving construct.* Unpublished doctoral dissertation, McGill University, Montreal, Canada.

Hale, G. A., Taylor, C., Bridgeman, B., Carson, J., Kroll, B., & Kantor, R. (1996). *A study of writing tasks assigned in academic degree programs*. Princeton, NJ: Educational Testing Service.

Hinkel, E. (2002). *Second language writers' text: Linguistic and rhetorical features*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hirvela, A. (2004). *Connecting reading and writing in second language writing instruction*. Ann Arbor, MI: The University of Michigan Press.

Howard, R. M. (1995). Plagiarisms, authorship, and the academic death penalty. *College English, 57*(7), 788-806.

Howard, R. M. (1999). The new abolitionism comes to plagiarism. In L. Buranen & A. M. Roy (Eds.), *Perspectives on plagiarism and intellectual property in a postmodern world* (pp. 87-95). NY: State University of New York Press.

Huff, K., Powers, D. E., Kantor, R. N., Mollaun, P., Nissan, S., & Schedl, M. (2008). Prototyping a new test. In C. A. Chapelle, M. K. Enright & J. M. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 187-225). New York: Routledge.

Hyland, K. (2006). *English for academic purposes: An advanced resource book*. London ; New York: Routledge.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342.

Kane, M. T. (2006). Validation. In R. Brennan, L. (Ed.), *Educational measurement*. Westport, CT: Praeger.

Kane, M. T., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17.

Keck, C. (2006). The use of paraphrase in summary writing: A comparison of L1 and L2 writers. *Journal of second Language Writing (2006), 15*(4), 261-278.

Lee, Y.-W., & Kantor, R. (2005). *Dependability of new ESL writing test scores: Evaluating prototype tasks and alternative rating schemes*. Princeton, NJ: Educational Testing Service.

Lichtman, M. (2006). *Qualitative research in education: A user's guide*. Thousand Oaks, CA: Sage publishers.

Lumley, T., & Brown, A. (2006). *Test taker response to integrated reading/writing tasks in TOEFL: Evidence from writers, texts, and raters*. Unpublished manuscript.

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. Mahwa, NJ: Lawrence Erlbaum Associates.

McNamara, T. F., & Roever, C. (2006). *Language testing: The social dimension*. Oxford, UK: Blackwell publisher.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). NY: Macmillan.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded source book*. Thousand Oaks, CA: Sage publications.

Oshima, A., & Hogue, A. (2005). *Writing academic English* (4th ed.). New York: Addison Wesley Longman.

Pearlman, M. (2008). Finalizing the test blueprint. In C. A. Chapelle, M. K. Enright & J. M. Jamieson (Eds.), *Building a validity argument for the test of English as foreign language* (pp. 227-258). NY: Routledge.

Pecorari, D. (2001). Plagiarism and international students: How the English-speaking university responds. In D. Belcher & A. Hirvela (Eds.), *Linking literacies: Perspectives on L2 reading-writing connections*. Ann Arbor, MI: The University of Michigan Press.

Pecorari, D. (2003). Good and original: Plagiarism and patchwriting in academic second-language writing. *Journal of Second Language Writing, 12*, 317-345.

Pennycook, A. (1996). Borrowing others' words: Text, ownership, memory and plagiarism. *TESOL Quarterly, 30*(2), 201-230.

Sawaki, Y., Stricker, L., & Oranje, A. (2008). *Factor structure of the TOEFL Internet-Based Test (iBT): Exploration in a Field Trial Sample*. Princeton, NJ: The Educational Testing Service.

Shi, L. (2004). Textual borrowing in second-language writing. *Written communication, 21*(2), 171-200.

Sutherland-Smith, W. (2005). Pandora's box: academic perceptions of student plagiarism in writing. *Journal of English for Academic Purposes, 4*, 83-95.

Swales, J. M., & Feak, C. B. (2004). *Academic writing for graduate students: Essential tasks and skills* (2nd ed.). Ann Arbor, MI: The University of Michigan Press.

The University of Melbourne. (2007a). *Postgraduate English language requirements*. Retrieved 22 November, 2007, from http://www.futurestudents.unimelb.edu.au/courses/pgenglishreq.html

The University of Melbourne. (2007b). *Undergraduate English language requirement*. Retrieved 22 November, 2007, from http://www.futurestudents.unimelb.edu.au/courses/ugenglishreq.html

Thompson, C. (2006). *Plagiarism or intertextuality? A study of the politics of knowledge, identity and textual ownership in undergraduate student writing.* Unpublished doctoral dissertation, University of Technology, Sydney, Sydney.

Thompson, G., & Yiyun, Y. (1991). Evaluation in the reporting verbs used in academic papers. *Applied Linguistics, 12*(4), 365-382.

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

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

Xi, X. (2008). Methods of test validation. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Language testing and assessment* (2nd ed., Vol. 7, pp. 177-196). New York: Springer

Appendix A: TOEFL iBT Test – Integrated writing rubrics (Educational Testing Service, 2005, p. 52)

Score	Task Description
5	A response at this level successfully selects the important information from the lecture and coherently and accurately presents this information in relation to the relevant information presented in the reading. The response is welt organized, and occasional language errors that are present do not result in inaccurate or imprecise presentation of content or connections.
4	A response at this level is generally good in selecting the important information from the lecture and in coherently and accurately presenting this information in relation to the relevant information in the reading, but it may have minor omission, inaccuracy, vagueness, or imprecision of some content from the lecture or in connection to points made in the reading. A response is also scored at this level if it has more frequent or noticeable minor language errors, as long as such usage and grammatical structures do not result in anything more than an occasional lapse of clarity or in the connection of ideas.
3	A response at this level contains some important information from the lecture and conveys some relevant connection to the reading, but it is marked by one or more of the following: ◎Although the overall response is definitely oriented to the task, it conveys only vague, global, unclear, or somewhat imprecise connection of the points made in the lecture to points made in the reading. ◎The response may omit one major key point made in the lecture. ◎Some key points made In the lecture or the reading, or connections between the two, may be incomplete, inaccurate, or imprecise. ◎Errors of usage and/or grammar may be more frequent or may result in noticeably vague expressions or obscured meanings in conveying ideas and connections.
2	A response at this level contains some relevant information from the lecture, but is marked by significant language difficulties or by significant omission or inaccuracy of important ideas from the lecture or in the connections between the lecture and the reading; a response at this level is marked by one or more of the following:

	◎The response significantly misrepresents or completely omits the overall connection between the lecture and the reading. ◎The response significantly omits or significantly misrepresents important points made in the lecture. ◎The response contains language errors or expressions that largely obscure connections or meaning at key junctures, or that would likely obscure understanding of key ideas for a reader not already familiar with the reading and the lecture.
1	A response at this level is marked by one or more of the following: ◎The response provides little or no meaningful or relevant coherent content from the lecture. ◎The language level of the response is so low that it is difficult to derive meaning.
0	A response at this level merely copies sentences from the reading, rejects the topic or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank.