Investigating the Summary Writing Performance of University Students in Taiwan

Shao Joyce Chin (金韶)
The Language Training and Testing Center
joycechin@lttc.ntu.edu.tw

This paper reports the initial findings of an ongoing investigation into the summary writing performance of university students in Taiwan. Forty-six students at the high-intermediate and advanced levels of proficiency participated in this study, in which they were asked to read two argumentative English texts and write a summary in English for each source text. Frameworks from Friend (2001), Rivard (2001), and Yu (2007) were adapted for data analysis. The 92 summaries were analyzed for the number and types of text propositions covered (i.e., main ideas, extraneous ideas and inaccurate statements) and then evaluated for the quality of writing (i.e., paraphrase and integration of ideas, rhetorical features, and language control). The performance data were triangulated with data obtained from post-task questionnaires and follow-up interviews.

The results showed that the high-intermediate level students, compared to their advanced-level counterparts, included fewer main ideas and more extraneous ideas in their summaries. The former group also appeared to be more constrained by lexical and grammatical deficiencies than the latter group did while writing summaries, resulting in more frequent use of verbatim replication and greater difficulty in reformulating and integrating ideas. Both groups needed more training on how to rearrange and reorganize the ideas extracted from a source text to compose a coherent summary text with adequate rhetorical fluency. The paper concludes with suggestions for teaching and further research.

Key words: summary writing, English for Academic Purposes (EAP), reading-writing connection, integrative task, analytical scoring, holistic scoring

INTRODUCTION

In the field of English for Academic Purposes (EAP), summary writing has been regarded as an integrative task that involves the interplay of two abilities—the ability to understand the main ideas and rhetorical organization of a text and the ability to compose a succinct and coherent restatement of the author's gist (Enright *et al*., 2000). Research on EAP training suggests that summary tasks can facilitate the holistic development of learners' English reading and writing abilities as well as their content area study skills (Radmacher & Latosi-Sawin, 1995; Friend, 2001). Nonetheless, previous studies (Cohen, 1994; Moore, 1997; Kim, 2001; Yang & Shih, 2003) have also shown that this integrative task is very challenging for ESL/EFL learners, as their deficiencies in reading and writing might lead to breakdowns at certain points in the process of summary production, ranging from identifying key words to integrating ideas into a coherent restatement.

Despite the above-mentioned pedagogical potentials and challenges of summary writing for EAP training, only a limited number of empirical studies have investigated the summary writing performance of university students in Taiwan. Some of these studies have explored the role of summarization in facilitating the development of writing ability. Hsu (2003), for instance, investigated the summary writing performance of 12 first-year English majors participating in web-based integrated reading and writing activities. She reported that summary writing helped the students conceive how to present the ideas in an organized manner and thus served as a "springboard" for them "to embark on composing tasks" (p. 18). Several other studies examined the effectiveness of summary writing as an instructional

strategy. Pan (2003), who experimented with two ways of using model essays in writing instruction, found that students who were instructed to summarize model essays regularly performed better in the post-course argumentative writing task than those instructed to read model essays extensively. In contrast to Pan's (2003) product-oriented approach, Chen (2002) analyzed students' think-aloud protocols to compare the strategies they used when completing different writing tasks. The results indicated that when writing a summary, the students employed relatively fewer writing strategies and high-level reading strategies than they did in writing a response essay.

These studies, though limited in number, have provided some glimpses of how university students in Taiwan deal with summary tasks and how this task type might foster reading and writing connection. However, in order to better understand how learners' developing reading and writing abilities affect their summary writing performances, there is an imperative need for empirical research which analyzes learners' written summaries in a more comprehensive manner, covering both the dimensions of content coverage and writing quality, and provides diagnostic information about their strengths and weaknesses in both dimensions.

Some recent empirical studies have provided insight into the development of a framework for analyzing learners' summaries and profiling their performances. Rivard (2001), through the collaborative effort of curriculum and language experts, identified ten "variables" for analyzing summaries. Four variables (*main ideas*, *secondary ideas*, *integration of ideas,* and *fidelity to the text*) focus on content-related issues; five variables (*organization, style, language usage, objectivity,* and *holistic writing score*) focus on language-related issues; and the last variable, *efficiency*, is related to both content and language. More recently, Yu (2007) developed a similar, but less complicated, evaluation framework incorporating two scoring schemes. The first scheme assigns points to each summary according to its adequacy and accuracy of content coverage. The second scheme evaluates the overall quality of the summaries based on a holistic scale that incorporates four key indicators: (1) *faithfulness to the source text*, (2) s*ummary and source text relationships*, (3) *conciseness and coherence,* (4) *ease of understanding*.

The above two researchers have identified some useful criteria for profiling students' summary writing performance. However, neither has provided an adequate treatment of learners' difficulties in paraphrasing and integrating ideas. Concerning these difficulties, Friend (2001), following the work done by Winograd (1984), developed a scheme for coding each sentence in a summary according to the amount of text from the original subsumed and the quality of paraphrase or integration. The scheme includes five categories: (1) *reproduction* (a single sentence rewritten), (2) *run-on* (two or more sentences poorly integrated), (3) *combination* (two or more sentence well integrated), (4) *low-level invention* (more than one paragraph poorly integrated) and (5) *high-level invention* (more than one paragraph well integrated). This coding scheme, nevertheless, does not include verbatim replication, a problem commonly observed in earlier empirical research on summary writing (Campbell, 1990; Johns & Mayes, 1990; Moore, 1997).

In the present study, a framework for analyzing and profiling written summaries was constructed on the insights provided by the above-mentioned studies. The framework includes three aspects: (1) a set of content-related criteria for analyzing the number and types of text propositions abstracted from the source text (i.e., main ideas, extraneous ideas and inaccurate statements), (2) a set of writing-related criteria (i.e., paraphrase and integration of ideas, rhetorical features, and language control), and (3) a holistic scale for evaluating the overall quality of the summaries, taking both content coverage and writing quality into consideration. The framework was employed to compare the summary writing performances of university students in Taiwan at two proficiency levels: the high-intermediate and

advanced levels. The study was designed to address the following research questions:

(1) In terms of content coverage (i.e., the number and types of text propositions abstracted from the source text), are there any significant differences in the summaries produced by Taiwanese university students at the high-intermediate and advanced levels of proficiency?
(2) In terms of the quality of writing (i.e., paraphrase and integration of ideas, rhetorical features, and language control), are there any significant differences in the summaries produced by Taiwanese university students at the high-intermediate and advanced levels of proficiency?
(3) When both content coverage and writing quality are considered, are there any significant differences in the summaries produced by Taiwanese university students at the high-intermediate and advanced levels of proficiency?

RESEARCH DESIGN

Participants

Forty-six EFL university students in Taiwan at the high-intermediate and advanced levels of proficiency participated in this study. The high-intermediate level students (N=23) were recruited from a pool of candidates who had scored between 80 and 90 on the LTTC-GEPT High-Intermediate Level Reading Test, and the advanced-level students (N=23) were recruited from a pool of candidates who had scored 90 or above on the LTTC-GEPT Advanced Level Reading Test[1]. The decision to recruit participants from those two pools was made on the basis of two considerations: (1) these two levels of tests are intended to measure the English abilities of university students in Taiwan, with the High-intermediate Level Test targeting non-English majors and the Advanced Level Test targeting English majors; (2) university students account for the largest proportion (approximately 60%) of the candidature of these two levels of tests.

Data Collection Procedures

The participants had to complete two summary writing tasks. For each task, they were given 60 minutes to read an argumentative text in English (approximately 800 words) and write a summary in English (approximately 180-200 words). A total of 92 summaries were produced by the 46 participants. While writing each summary, they had access to the source text but were not allowed to use dictionaries. Immediately after completing each task, they filled out a post-task questionnaire. In addition, before the first task session began, the participants were asked to fill out learner background questionnaires and then given a 10-minute orientation on summary writing task requirements. An overview of the data collection procedures is provided in Table 1.

To control text effects, counterbalancing procedures, illustrated in the shadowed parts in Table 1, were developed. Within each proficiency group, roughly half of the students were randomly assigned to Sequence 1, and the other half to Sequence 2.

[1] The cut-off scores for passing the LTTC-GEPT High-Intermediate and Advanced Reading Tests are both 80 (out of 120 points).

Table 1 Data collection procedures

Sequence 1 (11 HI* participants and 12 AD* participants)		**Sequence 2 (12 HI participants and 11 AD participants)**	
Procedures	Time (min.)	Procedures	Time (min.)
• Learner background questionnaire • Orientation on summary writing task requirements	15	• Learner background questionnaire • Orientation on summary writing task requirements	15
• **Summary Writing Task A** -summarizing the source text entitled "The Biopharming Controversy"	60	• **Summary Writing Task B** -summarizing the source text entitled "The Aquarium Controversy"	60
• Post-task questionnaire on Task A	10	• Post-task questionnaire on Task B	10
Break	10	*Break*	10
• **Summary Writing Task B** -summarizing the source text entitled "The Aquarium Controversy"	60	• **Summary Writing Task A** -summarizing the source text entitled "The Biopharming Controversy"	60
• Post-task questionnaire on Task B	10	• Post-task questionnaire on Task A	10

*The abbreviations "HI" and "AD" are used to refer to "high-intermediate level" and "advanced level", respectively.

In addition to collecting data from the two groups of university students using the above instruments, the researcher also conducted post-evaluation interviews with the raters to elicit their comments about the participants' summaries and the evaluation framework.

Materials and Instruments

1. Summary writing tasks

Alderson (2000) suggests that when designing a summary task, if the summarizers are informed of a purpose of summarization and provided with clearly-defined task requirements, the subjectivity of scoring can be reduced. In the present study, orientation sheets were provided before each task session to explain the requirements of the summary task. For each source argumentative text, the participants had to summarize the author's position on the issue, major arguments that oppose the author's position, and the author's rebuttals of the opposing arguments. They were also reminded that their summaries should be complete and coherent texts, restating the key arguments of the source texts and excluding personal opinions[2].

2. Source texts:

Two argumentative texts, one on the biopharming controversy and the other on the aquarium controversy, were developed as the source texts. The two texts were composed by a native speaker of English who is an experienced EFL writing instructor and had taught summary writing courses prior to this study. Efforts were made to maintain the comparability of the two source texts. Both texts were about 800 words in length and consisted of 6

[2] Summaries fall into different categories according to the classification schemes proposed by different researchers. In this study, the participants were required to write "reader-based" summaries (i.e., a stand-alone text written for an external audience, as defined in Hidi & Anderson, 1986), and as they had only to summarize the gist without adding comments, the summaries produced belong to the genre of "précis", following Bleck's (2003) classification (cited in Liu, 2004).

paragraphs. The syntactical complexity of the two texts was measured by the Flesh Reading Ease Score (39.1 and 39.2) and the Flesch-Kincaid Grade Level (both at Level 12), and the lexical diversity was measured by the type/token ratios (.47 and .43) and the percentages of words appearing in A. Coxhead's *Academic Word List* (8.15% and 5.38%)[3].

To determine the main ideas in each source text, five EFL writing instructors, consisting of three native speakers of English (hereafter abbreviated as NS) and two non-native speakers (hereafter abbreviated as NNS), were asked to write a summary of each source text within the same time limit as specified in Table 1. The eight most frequently occurring arguments (including the author's stated position, opposing arguments and the author's rebuttals of those arguments) were regarded as the "prespecified main ideas" of each source text. One of the source texts used in this study is provided in Appendix 1.

Data Analysis Procedures

As mentioned in the previous section, the participants' performances on the summary writing tasks were compared according to three sets of criteria: (1) content-related criteria for analyzing the number and types of text propositions abstracted, (2) writing-related criteria, and (3) a holistic scale for rating overall quality. The three sets of criteria were operationally defined as follows:

(1)-(3): Content-related criteria

(1) Main Ideas (MI): the coverage rate of "prespecified main ideas" identified by the five EFL writing instructors. The following unweighted partial credit system (cf. Yu, 2007) was adopted to assign Main Ideas scores of students' summaries:
- 2 points for a main idea adequately restated
- 1 point for a main idea included but inadequately restated or misplaced
- 0 point for no inclusion of the main idea

(The maximum score per summary was 16 points; Recurring statements of a main idea could be credited only once.)

(2) Extraneous Ideas (EI): the percentage of T-units[4] in a summary that should be excluded as unimportant ideas (with reference to the "prespecified main ideas"). A high percentage of extraneous ideas is an indicator of inadequate ability to summarize.

(3) Inaccurate Statements (IS): the percentage of T-units in a summary which were factual inaccuracies or over-generalizations. As in the case of extraneous ideas, a high percentage suggests unsatisfactory performance.

(4)-(6): Writing-related criteria

(4) Paraphrase & Integration (P&I): Friend's (2001) coding scheme was refined to assign points to each T-unit in a summary according to the amount of the text from the original subsumed and the quality of paraphrase or integration. The scheme included seven coding categories, as operationally defined in Table 2.

[3] Coxhead, A. (2000). A new academic word list. *TESOL Quarterly 34*(2): 213-238.
[4] A T-unit is generally defined as "one main clause with all subordinate clauses attached to it" (Hunt, 1965, p.20).

Table 2 Coding Scheme for Paraphrase & Integration (P&I)

P&I Coding Categories	Definition	Points
Verbatim Replication	Copying the whole sentence or more than five consecutive words from a sentence in the source text	0
Unsuccessful Reproduction	A single sentence poorly rewritten	1
Successful Reproduction	A single sentence well rewritten	4
Unsuccessful Combination	Two or more sentences (from one paragraph) poorly integrated	2
Successful Combination	Two or more sentences (from one paragraph) well integrated	5
Unsuccessful Integration	Ideas from more than one paragraph poorly integrated	3
Successful Integration	Ideas from more than one paragraph well integrated	6

To correct for the length of the summary, the sum of the points was divided by the *T* score of the number of T-units in the summary (excluding T-units judged as extraneous ideas).

(5) Rhetorical Features (RF): an analytical rating of the organization, coherence and cohesion of a summary using a six-band scale (1-6) adapted from the Michigan Writing Assessment Scoring Guide (Hamp-Lyons, 1991).

(6) Language Control (LC): an analytical rating of the lexical and grammatical usage in a summary using a six-band scale (1-6) adapted from the Michigan Writing Assessment Scoring Guide (Hamp-Lyons, 1991).

(7) Holistic Summary Quality (HSQ): Yu's (2007) holistic scale was adopted for rating the overall quality of the summaries, relating to both content and language. The scale contains six bands (A-F) and a three-level argumentation method (e.g., A-, A=, A+) was used, so the maximum score is 18.

The 92 summaries produced by the 46 students were typed and word processed; no attempts were made to correct spelling, punctuation, or grammatical mistakes. Two teams of raters were involved in the evaluation process. The first team of raters, including an EFL writing instructor (NS) and a research assistant (NNS), scored and coded the summaries according to the criteria of Main Ideas, Extraneous Ideas, Inaccurate Statements and Paraphrase & Integration. About 25% of the summaries were scored by both raters to establish an inter-rater reliability of .80 for each criterion. The remaining portion of the summaries were scored by one rater. The second team of raters, including two EFL writing instructors (one NS and one NNS), evaluated the overall quality of the summaries (HSQ scores) and provided analytical ratings on Rhetorical Features and Language Control. These two raters scored all of the 92 summaries. The inter-rater reliability (Spearman correlation coefficient) varied from .79 for the criterion of Holistic Summary Quality to .85 for the criterion of Language Control. Discrepancies were resolved by the researcher as the third rater[5].

[5] When the raters' band scores differed by one band, the scores were averaged. When they differed by two or more bands, the researcher determined the final score.

RESULTS AND DISCUSSION

Effects of proficiency level: an overview

The independent variable investigated in this study was the proficiency level (advanced vs. high-intermediate). Nine dependent variables were used to measure various aspects of the summaries written by university students at the two proficiency levels. The nine variables included the seven evaluation criteria operationally defined in the previous section, as well as the average numbers of words and T-units per summary. Statistical treatment of the data involved a series of independent-samples *t*-tests using "proficiency level" as the grouping variable. Table 3 presents the descriptive statistics by proficiency level, and Table 4 summarizes the results of independent-samples *t*-tests.

Table 3 Means and standard deviations by proficiency level

	Advanced Level (N=46)		High-Intermediate Level (N=46)	
Dependent Variables	Mean	SD	Mean	SD
Average no. of words per summary	179.800	22.469	183.630	18.452
Average no. of T-units per summary	14.890	3.025	15.540	2.641
Main Ideas	13.848	1.548	11.739	2.662
Extraneous Ideas	3.65%	4.91%	9.14%	8.41%
Inaccurate Statements	6.66%	9.14%	9.12%	8.86%
Paraphrase & Integration	.874	.184	.583	.197
Rhetorical Features	3.935	.523	3.315	.355
Language Control	3.924	.447	3.098	.442
Holistic Summary Quality	12.848	1.053	10.413	1.514

Table 4 *t*-test summary: effects of proficiency level on dependent variables characterizing the written summaries

Dependent Variables	*t*	Sig. (2-tailed)	η^2
Average no. of words per summary	-.893	.374	.009
Average no. of T-units per summary	-1.441	.153	.023
Main Ideas	4.644**	.000	.193
Extraneous Ideas	-3.815**	.000	.139
Inaccurate Statements	-1.308	.194	.019
Paraphrase & Integration	7.342**	.000	.375
Rhetorical Features	6.647**	.000	.329
Language Control	8.910**	.000	.469
Holistic Summary Quality	8.953**	.000	.471

**significant at $p < .01$

As can be seen in Tables 3 and 4, the summaries written by the high-intermediate level participants were longer and contained more T-units than those written by the advanced-level participants, although the differences were not statistically significant. Furthermore, gauging the participants' performances against the seven criteria in the evaluation framework, the advanced-level participants significantly outperformed ($p \leq .0005$) the high-intermediate level participants on six out of the seven criteria, except for the

criterion of Inaccurate Statements (p = .194). These results suggest that compared to their high-intermediate level counterparts, the advanced-level participants tended to write summaries which were slightly shorter, but generally more accurate, concise and coherent.

This initial finding awaits further fine-grained analyses. However, before continuing to comparison of the performances of the two proficiency groups in the aforementioned aspects of summary writing, we have to examine whether the use of two source texts, a moderator variable in this study, resulted in any significant performance differences or interactions between the variables, both of which might complicate the interpretation of the results.

Effects of source texts

Due to the limited number of participants who fit the sampling requirements, this study required each participant to write two summaries based on different source texts, with the intention of eliciting more data for analysis. As described in the previous section, several measures were adopted to ensure that the two source texts were comparable. In order to examine the effects of source texts on the dependent variables characterizing the written summaries, a series of 2 x 2 Mixed Design ANOVAs were conducted with "proficiency level" as the between-subjects variable and "source text" as the within-subjects variable. Only a moderate effect[6] of source texts was observed for the dependent variable of Extraneous Ideas (F = 4.739, p = .035, η^2 = .097).

As another measure of the effects of source texts, the post-task questionnaires included questions on how familiar the participants were with the topics of each source text. Three options were provided for them to indicate the degree of familiarity: (1) I am quite familiar with the topic; (2) I have heard or read about the topic once or twice; (3) I am not familiar with the topic. The percentages of participants who selected each option are presented in Table 5. A series of one-way ANOVAs were conducted to test the effects of topic familiarity on the dependent variables characterizing summary performances. No significant effects were found.

Table 5 Participants' familiarity with the topics of the source texts

	biopharming	aquarium
Quite familiar	2.17%	2.17%
Have heard/read about it once or twice	54.35%	63.04%
Not familiar	43.48%	34.78%

The above ANOVA results indicate that the use of different texts had little intervening effect on the participants' performances.

Profiling the summary writing performances of Taiwanese university students at the high-intermediate and advanced levels of proficiency

This section reports the findings from a series of comparative analyses on the performances of the two groups of participants, focusing on the content- and writing-related aspects of their summaries, as well as the overall quality of summarization.

[6] According to Cohen's (1988) criteria, an η^2 value ranged from .01 to .058 indicates a small effect; an η^2 value ranged from .059 to .138 indicates a moderate effect; and an η^2 value higher than .138 indicates a large effect.

1. Level differences in the content-related aspects of summaries

Among the three content-related dependent variables, statistically significant differences were found on two variables: Main Ideas (t = 4.644, $p \leq .0005$) and Extraneous Ideas (t = -3.815, $p \leq .0005$). The high-intermediate participants generally included fewer main ideas and more extraneous ideas in their summaries than did the advanced-level participants, as shown in the differences in their average Main Ideas Scores (11.739 and 13.848 for the high-intermediate and advanced groups, respectively) and in the percentages of T-units judged as extraneous ideas (.914 and .365). No significant level difference was detected on the dependent variable of Inaccurate Statements (t = -1.308, p = .194).

With respect to the variable of Main Ideas, an item analysis was conducted to further examine what types of main ideas were more likely to be omitted or misrepresented by the high-intermediate level participants. As mentioned in the research design section, eight "prespecified main ideas" were identified through textmapping procedures for each source text. These main ideas were classified into three types according to their argumentative functions:

- Type I: the author's overall position on the issue
- Type II: an opposing argument
- Type III: the author's rebuttal of an opposing argument

Based on the performance data on the high-intermediate group, mean item scores were computed for the eight "prespecified main ideas" in each source text, and the results are presented in Table 6.

Table 6 Mean item scores on Main Ideas obtained by high-intermediate participants

	MI 1	MI2	MI 3	MI 4	MI 5	MI 6	MI 7	MI 8
Text A (biopharming)	1.830	1.650	1.703	1.782	**1.265**	1.617	**1.096**	**.967**
Type	*I*	*II*	*III*	*II*	***III***	*II*	***III***	***III***
Text B (aquarium)	1.887	1.651	1.744	1.745	**.930**	**.893**	1.587	**1.131**
Type	*I*	*II*	*III*	*II*	***III***	***III***	*II*	***III***

An interesting finding derived from the data in Table 6 is that for both source texts, the three main ideas with the lowest mean item scores were all related to the author's rebuttals of the opposing arguments, indicating that the high-intermediate participants were more likely to omit or misrepresent the rebuttals than to do with the author's overall position and major opposing arguments. This finding corresponded to the results reported in Kintsch (1990) and Cohen (1994), that less proficient summary writers, who might be able to identify the topical information and overall thesis, often fail to differentiate finer levels of importance in the information gleaned, an indicator of inadequate competence in recognizing the hierarchical structure of text propositions.

Regarding the variable of Extraneous Ideas, Friend (2000) points out that novice summarizers tend to be distracted by what Garner & Gillingham (1989) labeled as "seductive details"—unimportant details that have the effect of enlivening an essay. In this study, several instances of this phenomenon were also found in the summaries written by the high-intermediate level participants.

2. Level differences in the writing-related aspects of summaries

2.1 Paraphrase and integration of ideas

As described in the section on research design, each T-unit in the participants' summaries was coded and scored according to the Coding Scheme for Paraphrase & Integration (P&I) summarized in Table 2. A total P&I score was computed for each summary to provide quantitative information about how well the ideas from the source text were paraphrased and integrated. An independent samples *t*-test was conducted to compare the mean P&I scores obtained by the high-intermediate and advanced groups. The results, as presented in Table 4, showed a significant difference by proficiency level (t = 7.342, $p \leq .0005$) with a fairly large magnitude of effect (η^2=.375).

To further compare how the two groups of university students processed and combined ideas from a source text, the percentages of different P&I types in each summary were computed[7]; the group means of the seven P&I types are listed in Table 7.

Table 7 Mean percentages of P&I types by proficiency level

P&I Types	Level	
	Advanced (N=46) Mean	High-Intermediate (N=46) Mean
Verbatim Replication	.047	.133
Unsuccessful Reproduction	.015	.064
Successful Reproduction	.301	.262
Unsuccessful Combination	.064	.106
Successful Combination	.363	.229
Unsuccessful Integration	.017	.029
Successful Integration	.116	.021

A 2 x 7 Mixed Design ANOVA was conducted with "proficiency level" as the between-subjects variable and "P&I type" as the within-subjects variable to test differences in the students' use of P&I types. The results showed significant effects of both the between-subjects variable (proficiency level, F = 14.474, $p \leq .0005$, η^2 = .139) and the within-subjects variable (P&I type, F = 86.07, $p \leq .0005$, η^2 = .489). A significant interaction effect between the two variables was also observed (F = 10.921, $p \leq .0005$, η^2 = .108). A further analysis of between-subjects simple effects pointed to significant performance differences between the two proficiency groups on the use of three P&I types: Verbatim Replication, Successful Combination and Successful Integration. These differences are discussed below.

First, with respect to the category of Verbatim Replication, the coding criterion, as described in Table 2, was whether a T-unit contained more than five consecutive words from a sentence in the source text. It was found that the high-intermediate level participants relied significantly more on copying than did their advanced-level counterparts (F = 8.80, p = .03). This problem was frequently reported in the summary research in the 1990's (Campbell, 1990;

[7] To compute the percentage of a particular P&I type in a summary, the number of T-units coded as that particular P&I type was counted and then divided by the total number of T-units in the summary, with the T-units judged as extraneous ideas excluded.

Johns & Mayes, 1990; Moore, 1997), and as the above result suggests, still deserves the continuing attention of EAP instructors.

Significant differences between the two proficiency groups were also observed on the percentages of T-units marked as Successful Combination (F = 21.56, $p \leq .0005$) and Successful Integration (F = 10.78, p = .001). The former involves combining ideas from sentences within a paragraph, and the latter involves across-paragraph integration of ideas. At both levels of processing, the advanced group outperformed the high-intermediate group. The following two pairs of samples demonstrate the qualitative differences in their performances.

Examples 1 & 2: Unsuccessful and successful combinations of ideas within a paragraph

Source text:

> If this technology works as intended, it could save an astounding amount of lives. Only 200 acres would be needed to produce enough vaccines to vaccinate every baby on Earth against hepatitis B, a disease which kills more than a million people annually. Even greater benefits could be had by the developing world. Because the plants could possibly be dried and shipped without the need of cold storage, the process of getting the vaccines to those in need could be drastically simplified. Furthermore, some scientists claim that plant-based vaccines which could be taken orally could be created, which would further reduce the cost of administering the vaccines.

Example 1 (Writer's code: HI-03)

> *Only smaller growing fields and less costs can those genetic modified plants easily conveyed and produced, unlike most of the corps in market.*

Example 2 (Writer's code: AD-02)

> *If the technology succeeds, it can help millions of people by manufacturing at low cost medical products that are easily processed & administered.*

Examples 3 & 4: Unsuccessful and successful integrations of ideas across paragraphs

Source text: Appendix 1

Example 3 (Writer's code: HI-18)

> *Although it is very economical and convenient* **[para. 2]**, *biopharming has the great risk of cross pollination, which lets a biopharm crop and a non-biopharm crop produce seeds and their crops containing doses of a drug but identical to regular ones.* **[para. 3]**

Example 4 (Writer's code: AD-01)

> *Biopharming is a new method to produce medical products that could help human efficiently* **[para. 2]**, *yet the author disapproves by stating that it causes cross-pollination, which might pollute other plants* **[para. 3]** *and affect animals and food chain* **[para. 4]**.

The writers of these four examples all attempted to generate one-sentence summaries of ideas either from within a paragraph or from more than one paragraph. However, in the case of high-intermediate level writers, these attempts were often undermined by inadequate lexical and grammatical competence, as demonstrated in Examples 1 and 3. In Example 1, the writer seemed to understand the global meaning of the paragraph, but syntactical errors (e.g., inaccurate passive forms), inappropriate word choice (e.g., *conveyed*) and collocation errors (e.g., *growing fields*) resulted in a restatement which cannot be understood with ease. In Example 3, the difficulties in reformulating ideas in his/her own words probably led to the writer's heavy use of verbatim replications, marked with underlining. These two examples highlight the possible effects of lexical and syntactical deficiencies on the summary writing performance of the Taiwanese university students at the high-intermediate level.

In comparison with the two high-intermediate examples, Examples 2 and 4, written by two advanced-level participants, demonstrate better paraphrasing and integration skills. In Example 2, all of the main points in the source paragraph have been incorporated into a well-written complex sentence and expressed predominantly in the summarizer's own words.

In Example 4, the main ideas from three paragraphs in the source text have been condensed to form the thesis statement of the summary. Errors in grammatical usage and collocation do exist (e.g., *help human efficiently*) but do not obscure or distort the meaning.

As shown in the above discussions, statistical analysis found significant differences between the two proficiency groups in the amount of verbatim replications and in their abilities to integrate ideas from different parts of the text. An analysis of summaries written by the high-intermediate level participants indicated that lexical and grammatical deficiencies might hamper their attempts to synthesize ideas. Interestingly, the post-task questionnaires revealed that the participants at both proficiency levels were aware of the constraint of limited vocabulary, particularly the difficulties in using synonyms, during the process of summarization, but few of them identified problems with syntactical structures as a source of difficulty. The discrepancy between performance and self-report data suggests an avenue for future research.

2.2 Rhetorical Features and Language Control

In addition to Paraphrase and Integration, this study investigated two more aspects of students' summaries, Rhetorical Features and Language Control. The participants' performances in these two aspects were measured by two separate 6-band analytical scales adapted from the Michigan Writing Assessment Scoring Guide (Hamp-Lyons, 1991). The scale for evaluating Rhetorical Features incorporated key indicators such as cohesion, coherence, organization, and rhetorical fluency, and the scale for rating Language Control focused on how well the grammatical structures and vocabulary chosen could express the ideas.

The results of independent samples *t*-tests, as summarized in Table 4, indicated that the advanced-level participants scored significantly higher than their high-intermediate counterparts on both Rhetorical Features ($t = 6.647$, $p \leq .0005$) and Language Control ($t = 8.910$, $p \leq .0005$). The values of η^2 (.329 and .469) indicated a fairly large magnitude of level effect on both dependent variables.

As shown in Table 3, the mean scores of the advanced group on both scales were close to 4, and the mean scores of the high-intermediate group on the two scales were either slightly above or close to 3. Based on the above patterns of score distribution, the following two summaries were chosen to illustrate typical summarization performances of the advanced and high-intermediate groups, as analyzed with reference to the scales for rating Rhetorical Features and Language Control.

Example 5
(Writer's code: HI-22; Score on Rhetorical Features: 3; Score on Language Control: 3)

> Nowadays, people had argued over the biopharming, which is a project of United States to produce medicine by farming corn, rice, and soybeans. If biopharming success, it could saved more lives in the developing world. *Otherwise*, every baby on Earth against hepatitis B will be saved because of this new technology. The scientists also claim that the cost of administering the medicine can decrease a lot.
>
> However, this new technology run the risks. The first significant risk is cross-pollination. If the testing plants contact with the normal plants, the seeds they produce might be got by a consumer and he (or she) might injest it without doctors' permission. The second risk is the animals who eat the testing corns might influence other animals through the food chain. The last but not the least risk is the effect of the micro-organisms, which could be killed by contacting the testing medicine.
>
> *[missing transition]*
>
> The US government have set strict regulations to avoid cross-pollination, but the US government still sometimes failed. The regulations can prevent human's behavior, but it can't

control animals' behavior from eating crops. Therefore, regulations cannot totally avoid any risks happening; it only can lower the risks.

Example 6
(Writer's code: AD-14; Score on Rhetorical Features: 4; Score on Language Control: 4)

Biopharming is a promising technology that uses genetically modified crops to produce the proteins we need for medical purposes. It seems like a blessing for those who need vaccines to vaccinate their descents.

Though it seems so wonderful, some problems are reported. First, cross-pollination makes it difficlt to control the dosage of vaccines in biopharm crops. Second, the biosystem may be disturbed after animals ingest these crops. Because of all these uncertainties of nature, the author opposes the idea of biopharming.

The pharmaceutical companies think up lots of ways to regulate the operations of biopharming, and feel that they've done enough. However, the failure in inspecting some test crops shows that the strict regulatory methods don't work. Furthermore, the tracking of crop development is conducted poorly, and can't give definitive answer for those who want to know what happened.

The mechanism of nature is too complex for us to regulate by man-made rules. The risk of hybrid crops and destruction of food chain can't be eliminated by the regulations. Our author think the drug-making companies should think seriously about the danger before they conduct the open-field testing.

In terms of rhetorical features, Example 5 received a score of 3, which was characterized in the rating scale as having "uncertain rhetorical control" (Hamp-Lyons, 1991: 274). This example reflects two problems frequently found in the summaries produced by the high-intermediate group. The first problem is unbalanced development of ideas. The writer used two long paragraphs (149 words) to summarize the first half of the source text and only one short paragraph (46 words) to abstract the second half, making the first two paragraphs of the summary too detailed and the last paragraph too vague. The second problem is illogical links among ideas, usually resulting from inaccurate use of connectors (e.g., "otherwise" in the first paragraph) or missing transitions (e.g., between the second and third paragraphs).

Example 6, written by an advanced-level participant, received a score of 4 on Rhetorical Features. Compared to the previous example, the summary was stronger in organization, coherence, and cohesion. Its organization was appropriate to the task and most of the textual elements were logically sequenced. Nevertheless, there were still instances of redundancy (e.g., "seems like a blessing" in the first paragraph and "seems so wonderful" in the second paragraph) and missing transitions across paragraphs (e.g., between the second and third paragraphs). Hamp-Lyons (1991: 274) considers occurrences of redundancy, repetition and missing transitions to be indicators of weak "rhetorical fluency." This example suggests that, in terms of rhetorical features, the advanced-level participants still had plenty of room for improvement.

With regard to the criterion of Language Control, the two raters unanimously gave Example 5 a band score of 3 and Example 6 a band score of 4. These two bands were distinct from each other in the amount of errors and the extent to which the errors interfered with the clarity of the summary. Example 5 contains more lexical and grammatical errors which obscure or alter the meanings of the source text. This result demonstrates again that lexical and grammatical deficiencies can affect the summary writing performance of high-intermediate university students in Taiwan.

3. Level differences in the overall quality of summaries

Thus far, we have compared the content- and writing-related aspects of the summaries written by the high-intermediate and advanced level university students participating in this study. This section will complete the comparative analysis by focusing on the overall quality

of their summaries.

As mentioned in the research design section, Yu's (2007) holistic scale, with six bands (A-F) and a maximum score of 18, was adopted for assigning scores on Holistic Summary Quality (HSQ). The statistical data, as summarized in Tables 3 and 4, once again revealed a significant difference by proficiency level. The mean HSQ score obtained by the advanced group was 12.848, and the mean HSQ score of the high-intermediate group was 10.413. *T*-test results indicated that the difference was statistically significant ($t = 8.953$, $p \leq .0005$) and the magnitude of level effect was fairly large ($\eta^2 = .471$).

Yu's (2007) holistic scoring scheme, as reviewed earlier in this paper, incorporates both the content and language dimensions of written summaries. The general guidelines are as follows (citing Yu, 2007, p.567):

- faithfulness to the source text, in terms of the percentages of right and wrong statements in the summary,
- the topological relationships between the summary and the source text, with emphasis on the use of the summarizer's own language and the integration and connections of the statements,
- the conciseness, coherence and logicality of communication of meaning,
- the overall difficulty that the raters might encounter.

Using the above general guidelines and the detailed rubrics provided by Yu (2007: 568-572), the raters gave an average HSQ score of 10 (equivalent to a C-) to Example 5 and a score of 13 (equivalent to a B-) to Example 6. According to the raters' comments, collected in follow-up interviews, both Examples 5 and 6 covered most of the prespecified main ideas, but had a couple of misinterpretations and omissions. Example 6 included more extraneous ideas than Example 5 did. However, it was the writer's inadequate control over rhetorical features and grammatical/lexical usage, as discussed on the previous page, that resulted in the lower score given to Example 5.

Another issue of interest is how the raters utilized the scoring rubrics to evaluate overall summary quality. During the post-evaluation interviews, the two raters were asked how they applied the criteria specified to assign a holistic score. One of the raters reported that she first compared the summary with the source text and the accompanying list of prespecified main ideas to check the accuracy of content coverage and spot instances of verbatim replication. After the first reading, she decided an initial score range. To finalize the rating, she then evaluated the summary for its writing quality. The other rater adopted a different approach. After the first reading, he gave an impressionistic score based on the writing quality of the summary, paying particular attention to its comprehensibility and clarity. Then he read the summary again to examine its faithfulness to the source text and the extent to which ideas were reformulated and integrated.

Although these two raters approached the evaluation task in distinctive ways, both of them endorsed the idea of analyzing summaries based on a framework covering both the dimensions of content coverage and writing quality. They noted that diagnostic information about the learners' strengths and weaknesses in different aspects of summary writing can help course designers and instructors to develop materials and pedagogies tailored to the needs of their target students.

CONCLUSION

This study was designed to compare the summary writing performances of university students in Taiwan at the high-intermediate and advanced levels of proficiency. The results of statistical analysis indicated that the summaries written by the advanced-level university

students participating in this study were generally more accurate, concise, and coherent than those written by the participants at the high-intermediate level. More fine-grained analyses showed the following performance characteristics:

(1) In the dimension of content coverage, the high-intermediate level students, compared to the advanced-level students, included fewer main ideas and more extraneous ideas in their summaries. Although they were able to identify the topical information and overall thesis of the source text, they experienced difficulty in reconstructing a complete hierarchical structure of the ideas in the source text.

(2) In the dimension of writing quality, the high-intermediate level students resorted to verbatim replications more frequently and were less skillful at reformulating and integrating ideas, as compared to their advanced-level counterparts. However, both groups needed more training on how to rearrange and reorganize the ideas extracted to compose a coherent summary text with adequate rhetorical fluency.

(3) The high-intermediate level students were more likely to be constrained by deficiencies in lexical and grammatical competence, which affected the accurate conveyance of main ideas, the coherence and organization of summary texts, and thus the results of overall evaluation.

Based on these performance profiles, the following pedagogical suggestions should be considered by those who intend to incorporate summary writing into EAP training for university students in Taiwan. With regard to high-intermediate level students, the instructors may wish to provide continued guidance on the use of synonyms and other paraphrasing strategies. Students at this level also need further practice in combining sentences within, as well as across, paragraphs, and the importance of syntactical skills should not be neglected. Furthermore, students at both the high-intermediate and advanced levels need to work on recognizing the logical and rhetorical relationships of ideas in a source text and using appropriate connectors or transitions to represent the macrostructure established in that text.

Finally, it is necessary to note the limitations of this study and suggest avenues for future research. First, since a limited number of university students at only two proficiency levels were sampled, the generalizability of the results is contingent upon validation with more empirical evidence. It is suggested that further comparative analyses of this nature should be conducted with a larger sample of students at a wider range of proficiency levels. Second, this study focused only on the products of summary writing, To gain further insight into the processes and strategies involved in summarization, introspective data should be collected through mentalistic measures (e.g., concurrent or retrospective verbal reports). It is hoped that these lines of research can help generate more research-based materials and pedagogies to enhance the summarization performance of university students in Taiwan.

REFERENCES

Alderson, J. C. (2000). *Assessing reading.* Cambridge: Cambridge University Press.

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

Chen, Y. C. (2002). *Writing from Sources: Summary and response writing by four EFL graduate students.* Unpublished master's thesis. Providence University, Taiwan.

Cohen, J. (1988). *Statistical power analysis for the behavior science.* Hillsdale, NJ: Lawrence Erlbaum.

Cohen, A. D. (1994). English for academic purposes in Brazil: The use of summary tasks. In C. Hill & K. Parry (Eds.) *From testing to assessment: English as an international language* (pp. 174-204). London: Longman.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly 34*(2): 213-238.

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. A. (2000). *TOEFL 2000 reading framework: A working paper.* Princeton, NJ: Educational Testing Service.

Friend, R. (2000). Teaching summarization as a content area reading strategy. *Journal of Adolescent and Adult Literacy. 44*(4), 320-329.

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

Garner, R., & Gillingham, M.G. (1989). Effects of "seductive details" on macroprocessing and microprocessing in adults and children. *Cognition and Instruction, 6*, 41-57.

Hamp-Lyons, L. (1991). *Assessing Second Language Writing in Academic Contexts.* Norwood, NJ: Ablex Publishing Corporation.

Hidi, S., & Anderson, V. (1986). Producing written summaries: Task demands, cognitive operations, and implications for instruction. *Review of Educational Research, 56*(4), 473-493.

Hsu, H. C. (2003). *A case study of the process of web-based reading and summary writing for EFL college novice writers in Taiwan.* Unpublished doctoral dissertation. National Taiwan Normal University, Taiwan.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels.* Urbana-Champaign, IL: National Council of Teachers of English.

Johns, A.M. & Mayes, P. (1990). An analysis of summary protocols of university ESL students. *Applied Linguistics 11*, 253-271.

Kim, S. A. (2001). Characteristics of EFL readers' summary writing: A study with Korean university students. *Foreign Language Annals, 34*(6), 569-581.

Kintsch, E. (1990). Macroprocesses and microprocesses in the development of summarization skill. *Cognition and Instruction, 7*(3), 161-195.

Liu, S. H. (2004). *A study on the effects of English summary writing instruction on Taiwanese senior high school students' reading and writing abilities.* Unpublished master's thesis. National Taiwan Normal University, Taiwan.

Moore, T. (1997). From text to note: cultural variation in summarization practices. *Prospect, 12*, 54-63.

Pan, Y. C. (2003). *A study of the effects of reading and summarizing model essays on EFL writing*. Unpublished master's thesis. National Tsing Hua University, Taiwan.

Radmacher, S. A., & Latosi-Sawin, E. (1995). Summary writing: A tool to improve student comprehension and writing in psychology. *Teaching of Psychology, 22*(2), 113-115.

Rivard, L. P. (2001). Summary writing: A multi-grade study of French-immersion and Francophone secondary students. Language, *Culture and Curriculum, 14*, 171-186.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* New York: Academic Press.

Winograd, P. N. (1984). Strategic difficulties in summarizing texts. *Reading Research Quarterly*, 19, 404-425.

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

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

Yu, G. (2005). *Towards a model of using summarization tasks as a measure of reading comprehension.* Unpublished PhD thesis, University of Bristol, Bristol.

Yu, G. (2007). Students' voices in the evaluation of their written summaries: Empowerment and democracy for test takers? *Language Testing, 24*(4): 539-572.

Appendix 1: Source text (The Biopharming Controversy)

Biopharming is the genetic modification of plants to produce proteins which can be extracted and turned into medicines and vaccines targeted at a host of diseases and conditions which plague mankind today. In the United States, there are hundreds of open-field test crops of corn, rice, soybeans, which promise to yield hundreds of new medical products. While the technology looks promising, closer inspection reveals that not only are these promises far from guaranteed, but biopharming could have disturbing consequences.

If this technology works as intended, it could save an astounding amount of lives. Only 200 acres would be needed to produce enough vaccines to vaccinate every baby on Earth against hepatitis B, a disease which kills more than a million people annually. Even greater benefits could be had by the developing world. Because the plants could possibly be dried and shipped without the need of cold storage, the process of getting the vaccines to those in need could be drastically simplified. Furthermore, some scientists claim that plant-based vaccines which could be taken orally could be created, which would further reduce the cost of administering the vaccines.

Unfortunately, with these potential benefits are great risks, chief among them is the risk of cross-pollination. Cross-pollination occurs when pollen from the flower of one plant comes into contact with another to produce seeds which possess characteristics from both parent plants. The danger is that if cross-pollination occurred between a biopharm crop and a non-biopharm crop, the seeds produced might carry doses of a drug. Those seeds would be identical to regular seeds and would produce grains, fruits, and vegetables which would not show any sign that they contained a drug. Neither the farmer who plants the seeds nor the consumer who later ingests the produce will detect the difference. As a result, the consumer of these plants could be ingesting medicine that should only be administered in controlled amounts under a doctor's orders. Though pharmaceutical companies claim the quantity would be too small to affect a human, they have not provided evidence supporting this.

Even if a human may be insusceptible, what about smaller animals which ingest these crops? The effects could certainly be greater for animals the size of a rabbit or bird. It is possible that these drugs could build up in the liver of these animals, and later these animals might be eaten by humans, providing a mega-dose of the drug. It is also possible that the dose might kill these animals. If enough animals were affected in a given species, the food chain itself could be disturbed. Other at-risk life forms are micro-organisms, which could be harmed or even killed by even the smallest exposure. Though micro-organisms are small, they play a mighty role in the life of every plant and animal on Earth. Not even the simplest bodily function could occur without them. No one knows with certainty whether micro-organisms would be affected, but if so, the effects could be catastrophic.

Proponents of biopharming feel controlling the risks is within our means. They presume that we can prevent cross-pollination with strict regulations. Indeed, many strict regulations have been created by the US government, covering all aspects of biopharming. The Animal and Plant Health Inspection Service (APHIS), under the jurisdiction of the United States Department of Agriculture (USDA), has been charged with enforcing them. One of the regulations is that every biopharm test crop must be inspected by APHIS. Nevertheless, in December 2005, the USDA released a report that revealed APHIS had not only failed to inspect some test crops but were not even aware of the locations of others. Furthermore, though the tracking of each crop through to their harvest is also required, APHIS did not make any inspections beyond an initial one on several crops. As a result, APHIS can not definitively say what happened to them. Pharmaceutical companies, on the other hand, claim that despite the APHIS' failures, they have complied with the regulations. However, records indicate otherwise, as several companies have already been cited for allowing cross-pollinations.

In fact, cross-pollination as a result of poorly enforced regulations is not the central issue for critics of biopharming. For them, what the argument boils down to is that no man-made regulation can possibly account for the unpredictability of nature. Regulations cannot prevent birds, animals, insects, or the wind from spreading pollen and seeds. They cannot prevent birds or animals from eating seeds. They cannot ensure that seeds are not left behind to sprout long after experiments are over or prevent seeds from falling off trucks during transport. Regulations can not eliminate these risks. For this reason alone, the biopharm industry needs to seriously rethink the open-field testing of biopharm crops.