

Task Type and Prompt Effect on Test Performance: A Focus on IELTS Academic Writing Tasks

A. Ahmadi *
Associate Professor, TEFL
Shiraz University
email: arahmadi@shirazu.ac.ir

S. Mansoordehghan
M.A., TEFL
Islamic Azad University of Hashtgerd
email: dehghan242@gmail.com

Abstract

Recent versions of international high-stakes tests like TOEFL and IELTS have made use of integrated tasks in addition to the traditional independent tasks in a claim to provide a more realistic estimation of the test takers' language abilities. The present study aimed to investigate how test takers' performance may differ on such tasks. As such, the test takers' performance was compared on IELTS Academic Writing Tasks 1 and 2. Whereas Task 1 is an integrated task which calls for graphic interpretation and description, Task 2 is an independent task of writing an argumentative essay. Furthermore, the study also aimed to investigate the effect of writing prompts on the test takers' performance on such tasks. The study adopted a quasi-experimental design in the form of posttest-only group. Fifty six Iranian EFL learners at Shiraz University were selected based on their availability. After receiving instruction in a semester on how to do the two tasks, they received 4 writing tasks (2 versions of each task with different prompts). The data collected were analyzed through two-way repeated-measures ANOVA. The results indicated that task type did not have a significant effect on the test takers' writing performance; that is, there were no significant differences between the participants' performance on the independent and integrated writing tasks. Furthermore, the effect of prompts was only found to be significant on the participants' performance on task 2 (independent task). The findings provide evidence for higher consistency of scores obtained from different versions of the integrated task.

Keywords: IELTS, writing tasks, task type effect, writing prompts, Iranian EFL learners

Received: 11/17/2013 Accepted: 10/15/2014

*Corresponding author

1. Background to the Study

Since 1970s, more and more researchers have focused their attention on assessing the academic writing abilities of university students with the introduction of direct writing tasks (Hamp-Lyons, 1991). Furthermore, “university administrators ask students to provide indicators of their writing ability to help in making either admission or placement decisions” (Gebril, 2009, p. 508). In addition to the context-specific writing tests which are used by universities and institutes in this regard, standardized tests like TOEFL and IELTS play a significant role in admission or placement decisions. This has drawn lots of attention to such tests; and many researchers (e.g., Cho, Rijmen, & Novak, 2013; Moore & Morton, 2007; Read & Hayes, 2003) have tried to investigate the validity of such tests in terms of the language tasks used in the tests, the scores obtained and their interpretations, the decisions made on the basis of such scores, and the consequences of such decisions.

Among the studies focusing on the writing section of IELTS, investigations have been conducted on determining the appropriate band scores for admission into programs (Golder, Reeder, & Fleming, 2011; Green, 2005), comparing IELTS writing tasks with the university writing (Moore & Morton, 1999, 2005), investigating rater variation in scoring (Gao & Brennan, 2001; Lee & Kantor, 2005; Schoonen, 2005), examining IELTS as an indicator of written proficiency levels (Ellis, Chong, & Choy, 2013), elaborating on IELTS gain scores (Brown, 1998; Elder, & O’Loughlin, 2003; Read & Hayes, 2003), and finally studying the washback effect of IELTS and the impact of preparation programs on candidates’ performance (Green, 2007; Rao, McPherson, Chand, & Khan, 2003).

Some of the studies in this regard have pinpointed the problems of conventional independent writing tasks in which test takers are provided with a topic and are asked to write about it within a specific time limit. It is mentioned that independent writing tasks are decontextualized as they do not let the test takers benefit from other sources while constructing their texts (Hamp-Lyons & Kroll, 1996); they require topic familiarity and therefore, some test takers may not be able to fully show their writing ability due to the lack of topic familiarity (Gebril, 2006) and finally such tasks may underestimate test takers’ writing competence (Hamp-Lyons & Kroll, 1996). Such criticisms have led to the emergence of integrated writing tasks in recent versions of TOEFL iBT, IELTS, etc. In contrast to the independent writing tasks in which test takers have to rely on their knowledge of topic to

produce a text, in integrated writing tasks, they can benefit from other sources in constructing their texts. This will need integration of several skills at the same time. Therefore, they may listen to a tape or read a text and then write a text based on the information they have received and their background knowledge. It is stated that integrated tasks reflect authentic academic tasks (Wiegler, 2004), and to improve the strategic competence as test takers need to apply different strategies to complete such tasks (Plakans, 2009).

However, a few empirical studies have compared independent and integrated writing tasks to see how test takers' performance may really differ on such tasks and whether the inclusion of such tasks together will better represent one's writing ability than the traditional independent tasks alone. Of such comparative studies, mention could be made of Moore and Morton (2005) who compared the IELTS writing Task 2 with a corpus of 155 assignment tasks collected at two Australian universities. The results indicated that although there existed some similarities between this type of writing (Task 2), and the predominant genre of university study (the essay), there were some important differences between the two as well. They concluded that the type of writing that IELTS Task 2 elicits has more in common with nonacademic genres and as such cannot be considered appropriate in eliciting a sample of university writing.

In another study, Ellis et al. (2013) examined the writing proficiency of graduating secondary student teachers at the National Institute of Education (NIE), Singapore. The writing proficiency was measured through IELTS writing scores. The results indicated that the participants scored lowest on the writing skill than on the other skills, meaning that writing was the weakest skill of such teachers. The overall writing scores ranged from 3.5 to 8.5, indicating that such teachers were more heterogeneous on this skill than other language skills. Although the overall mean score of writing was 7.2; that is, a little higher than the minimum score of 7 (good user) which is required by NIE for foreign language teachers, about 20% could not reach this criterion as they had a mean score of 6.5 or lower. Furthermore, the mean score of 6.2 on the writing Task 1 indicated that they were better at argumentative writing, whereas they had more problems describing a visual prompt like a table or a graph.

But as mentioned, the literature suffers from empirical studies comparing independent and integrated writing tasks. This research gap is greatly felt when we focus on how the type of prompt may affect test takers'

performance on IELTS Academic Writing Tasks 1 and 2. To the best of the researchers' knowledge, no study has been conducted so far to specifically focus on the effect of prompt on test performance on task 1 or 2 of IELTS. The present study was, therefore, an attempt in this regard to contribute to the literature concerning the comparison of independent and integrated writing tasks. The participants' test performance was compared on IELTS Academic Writing Tasks 1 and 2. Furthermore, the study aimed to investigate the effect of prompts on test performance. That is, because the two tasks are of different nature (integrated vs independent), seeing how test takers' performance on each task is influenced by different prompts would provide more insights into the dependability of the results and the interpretations made based on the scores. The study specifically sought answers to the following research questions:

1. How does test takers' performance differ on IELTS Academic Writing Tasks 1 and 2?
2. Does the type of prompt affect performance on IELTS Academic Writing Tasks 1 and 2?

2. Method

2.1 Research design

This study adopted a quasi-experimental design in the form of posttest-only paradigm. In this design, the participants receive treatment, and then they are given a posttest to see how they perform on the test. The focus of such a design is usually performance rather than development (Mackey & Gass, 2005).

2.2 Participants

The participants of the present study were 56 senior students majoring in English Language and Literature at Shiraz University who were selected based on their availability. They all had enrolled in an essay writing course. This course was the continuation of another writing course (Paragraph Development) they had successfully passed in the previous semester. The purpose of the course was to help them become familiar with different types of writing such as expository essays, argumentative essays, and online writing. The participants were attending the course in two classes taught by the same instructor.

2.3. Materials

IELTS academic writing tasks 1 and 2

Two writing tasks (IELTS Academic Writing Task 1 & 2) were used in this study to measure the students' writing ability after a period of instruction. To study the effect of prompt, each writing task appeared in two different prompts. Therefore, overall four writing tasks were given to the participants. The writing tasks were selected from the book *Academic Writing Practice for IELTS* (McCarter, 2002). As mentioned before, the purpose of Task 1 in the Academic Writing Module is to replicate an academic writing by asking test takers to base their writing on a source which is presented in the form of a bar graph, pie chart, table, or diagram. They should write at least 150 words in 20 min. The task demands graphic interpretation and description and as such can be considered an integrated task of writing. Task 2 is longer than Task 1 (a minimum of 250 words) and takes more time (40 min). It also carries more weight in scoring (Uysal, 2010; Weigle, 2002). It requires the test takers to write an essay in response to an argument, a problem, or a proposition. Test takers should use factual information, logical explanations, evaluative judgments, and personal examples to support their opinions. As such, this task is basically an argumentative type of writing.

2.4 Treatment

The participants received instruction and training on how to write expository essays, argumentative essays, and IELTS Writing Tasks 1 & 2 during the academic semester. The essay writing class was held one session (90 min) a week during a full academic semester of 16 weeks. Each session, the instructor gave the participants explicit instructions on one of the above types of writing. He also taught them certain strategies needed for successful writing. The instructor also focused on sample models of writing, and together with the participants tried to analyze the samples. After the instruction and analysis of the models, the participants were supposed to write on a variety of topics. The class was mostly an interactive one, focusing on process-writing. The participants received feedback on their individual as well as pair and group writings. This feedback was provided to them both during the writing task while they were busy writing and also at the end of the writing task when the final drafts were collected. That is, the instructor collected the final drafts, corrected them, and provided the participants with feedback in the next session. Some writing practice was also done outside class; that is, in addition to the practice they had in their

regular class hours, they also had some writing assignments to be done at home. They received feedback on these assignments as well. For the purpose of the study, the writing tasks were basically taken from the book *Academic Writing Practice for IELTS* (McCarter, 2002). However, a few of them were taken from other similar sources as well.

2.5 Data collection procedure

The data were collected after the students had received explicit instruction and training on Writing Tasks 1 and 2 of IELTS. The four writing tasks were given in two different sessions. Counter balancing was used to control for the effect of ordering. In the first session, the students were first given Task 2 which asked them to write an argumentative essay on whether spending huge amounts of money on projects investigating the possibility of living in other planets is logical. Following the time limitation given in IELTS, they were given 40 min to write an essay of at least 250 words. Then, all the papers were collected, and the second test (Writing Task 1) was given to them. They were supposed to describe a table of statistics on the percentage of pupils entering higher education from five secondary schools between 1995 and 2000. They had 20 min to finish this task. In the second session, the same procedure was followed except for the fact that this time the tasks were given in a different order; at first Task 1 was given and then Task 2. Task 1 focused on describing a table of statistics indicating the percentage of rooms occupied in six hotels during May to September between 1985 and 2000, and Task 2 was about the benefits and dangers of using the Internet.

2.6 Scoring procedure

All the tasks were scored based on the analytic rating scale used in scoring IELTS writing tasks. This scale reports band scores between 1 (nonuser) to 9 (expert user). The ratings were done by an experienced instructor with expertise in language testing who had already taught writing courses such as English grammar, letter writing, paragraph development, and essay writing at B.A. level, and advanced academic writing at M.A. level for several years.

2.7 Intrarater reliability

All the writing samples were rated within two days after their collection. About 20% of the papers were also randomly selected and rerated by the researcher (instructor) after an interval of two weeks. To eliminate any

effect on the ratings of the rater-participant interaction, since the instructor was completely familiar with the participants, all the papers were rated anonymously in the two ratings. The intra-rater reliability was then checked which turned out to be acceptable ($r = 0.87$).

2.8 Data analysis

A two-way repeated-measures ANOVA was conducted to compare the scores on IELTS Writing Tasks 1 and 2 across the two versions of each task. The purpose was to see whether the participants' performance on the two tasks and on different versions of the same task (with different prompts) differed significantly or not. All the assumptions for using ANOVA including normality (based on kurtosis and skewness values and the normality test) and sphericity were checked to make sure that the use of parametric statistics was plausible. Sphericity assumption is automatically met when a variable has only 2 levels (Tabachnick & Fidell, 2007, p. 46), which was the case in this study.

3. Results

Table 1 summarizes the descriptive statistics of performance on the four tasks.

Table 1 . Descriptive statistics of the performance on IELTS academic writing tasks 1 and 2

Task Type	<i>N</i>	<i>Range</i>	<i>Min.</i>	<i>Max.</i>	<i>M</i>	<i>SD</i>
GD1	56	4	4	8	6.45	.76
GD2	56	3.5	4.5	8	6.50	.78
AE1	56	6	2	8	6.20	1.59
AE2	56	3.5	4.5	8	6.81	.81

GD = Graph Description, AE = Argumentative Essay

As depicted, the mean scores of performance on the two versions of Task 1 are very similar. Furthermore, it is illustrated that the performance on the two argumentative essays (Task 2) is different with the mean score on the first task being lower. It is also interesting that these two tasks embrace higher variation of scores with the first version indicating the highest variation among all the tasks ($SD = 1.62$).

To see whether the differences depicted in Table 1 are significant or not, a two-way repeated-measures ANOVA was employed. The results (Table 2) indicated no significant effect for the task type, meaning that the test takers' performance on the two tasks did not differ significantly. However, significant differences were found for the prompt effect and the interaction effect of task and prompt with both indicating large effect sizes based on Cohen (1988).

Table 2. Multivariate test results for task version

	Value	F	Hypothesis df	Error df	Sig.	Partial eta squared
Task Type Wilks' Lambda	.999	.06	1.000	55	.807	.001
Task version (prompt effect)	.875	7.83	1.000	55	.007	.125
Task type*prompt	.860	8.925	1.000	55	.004	.140

To see where exactly the significant differences lay, Bonferroni post hoc tests were employed. The results are depicted in Table 3.

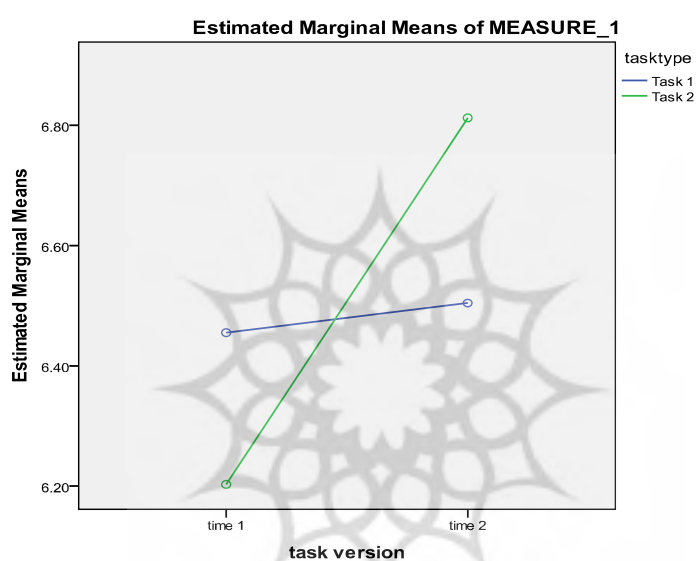
Table 3. Pairwise comparisons of tasks based on Bonferroni

Tasks	Mean Difference	Std. Error	Sig.	Lower Bound	Upper Bound
GD1-GD2	-.049	.088	1.000	-.289	.191
GD1-AE1	.253	.188	1.000	-.262	.767
GD1-AE2	-.357	.094	.002	-.616	-.099
GD2-AE1	.302	.211	.945	-.275	.878
GD2-AE2	-.308	.089	.006	-.550	-.066
AE1-AE2	-.610	.194	.016	-1.141	-.079

Table 3 indicates that the significant differences appear between the second version of the argumentative essay (Task 2) on the one hand and all the other tasks on the other. The results depicted could be more specifically stated as follows: (a) there existed no significant difference between different versions of Task1 (integrated task), meaning that the participants had the same performance on Task 1 regardless of different prompts; (b) there existed a significant difference between different versions of Task 2,

meaning that the participants' performance differed significantly from one version of Task 2 to the other because of the prompt used; and (c) both versions of task 1 differed significantly from one of the versions of task 2. This indicates the greater impact of prompt than the task type. The results are more tangibly depicted in Figure 1.

Figure 1. Interaction effect of task type and task version



4. Discussion

The first research question of the study concerned the test takers' performance on the two academic writing tasks of IELTS. The results indicated that task type did not have a significant effect on the test takers' writing performance. This finding is interesting as the two tasks are not the same in terms of language functions they require. Whereas Task 1 includes, at least, two steps of graph interpretation and description, Task 2 calls for a kind of evaluative language and argumentative judgment. In fact, Task 1 is an integrated task of writing, whereas Task 2 is an independent writing task which has traditionally been used in many L2 writing exams. Such independent tasks have been criticized by many researchers (e.g., Cho, 2003; Gebril & Plakans, 2009; Hamp-Lyons & Kroll, 1996; Plakans, 2007; Weigle, 2002) for their inadequate measuring of the academic writing construct and also for their lack of relevance to the real-life contexts. For example, Moore and Morton (2005) criticizing IELTS Task 2 state that:

Preparation for the IELTS writing test (Task 2) may not give students an entirely accurate view of the nature of academic argumentation, especially concerning the issue of what constitutes appropriate evidence in a piece of writing. In the IELTS test, students learn that it is sufficient to base their assertions on their own ideas, knowledge and experience. In the university context- where valid evidence is usually seen as the findings of research or the authoritative pronouncements of disciplinary scholars—a student who relies exclusively on prior knowledge will usually be criticized for being anecdotal and for not having read adequately for the task. (p.54)

Such criticisms have recently led to the use of integrated tasks in standardized tests including IELTS. The claim is that integrated tasks have more in common with the academic writing. For example, Moore and Morton (1999, 2005) found that IELTS Writing Task 1 was more in line with the academic genre, whereas Task 2 was more in line with nonacademic genres. They, therefore, suggested that integrated reading-to-write tasks be included in IELTS.

It is stated that integrated tasks are similar to academic tasks in the sense that writing is based on a source and this may reduce the bias in the test and increase test fairness as the test takers who come to the testing situation with different background knowledge start their writing from the same base source (Plakans, 2007). Although IELTS Academic Writing Task 1 may not seem to include the same integration that exists in reading-to-write tasks that, for example, you cannot benefit from citations in such tasks; it is, however, considered an integrated task because it “requires the dual abilities of comprehension of graph input and transformation of visual information into written discourse” (Yang, 2012, p. 174).

The results of the present study indicate that the differences mentioned in some studies in the literature between integrated and independent tasks in terms of eliciting different behaviors and writing performance (e.g., Guo, 2011), are not necessarily determined by the scores obtained on the two tasks. Some studies have indicated that test takers will have better performance on integrated writing tasks (e.g., Grabe, 2001; Spack, 1993). Unlike such studies, the current study led to the results that are in line with Gebril’s (2006) indicating a high correlation between the scores on integrated and independent writing tasks. This can mean that the two tasks

are tapping the same construct or relevant aspects of the same construct; that is, academic writing. As such, using both tasks together instead of reliance on only one of them would benefit the test takers more. Similarly, Gebril (2009) found that the scores obtained on integrated tasks were as reliable as those obtained on independent writing tasks. Furthermore, the results indicated that score generalizability was very low when only one writing task was used. However, it should be noted that none of these studies has specifically focused on IELTS academic writing tasks. They have compared a certain type of an integrated task with an independent task. As an integrated task may come in different variations, caution is needed in comparing the results of such studies as the agreement or disagreement among the findings may be due to the different variations of the task used.

The second research question focused on the effect of prompt on test performance. The mean scores found for the test takers' performance on the two versions of task 1 (integrated task) with different prompts were very similar with no significant difference. However, a significant variation was found between the scores obtained on the two versions of Task 2 (independent task) indicating the prompt effect. Out of the four tasks, the highest and lowest mean scores were related to Task 2, which means that this task was subject to more variation. In contrast, the mean scores on the two versions of Task 1 were very closely related. This can indicate the high reliability of Task 1, which is an integrated task, in terms of score consistency across different versions of the task. Therefore, the scores obtained from IELTS Academic Writing Task 1 can be considered comparable across different versions of this task. In other words, test takers' performance is not significantly influenced by different prompts of different versions of this task.

However, it should be mentioned that IELTS Task 1 appears in several different formats (describing a table of statistics, bar-graph, line-graph, pie chart, a combination of graphs, etc.). The present study only focused on table description and found that this task led to consistent results. It is, however, possible that descriptions of other types of graphs are not consistent as different graphs could be of different difficulty level for comprehension and interpretation. IELTS administrators state that the tasks used in this test are pretested to ensure that they are appropriate in terms of content and level of difficulty (Uysal, 2010). But the results of such attempts could not be foolproof. For example, O'Loughlin and Wigglesworth (2003) found differences among tasks in terms of the language used. Simpler tasks

with less information led to more complex and better language production on the part of test takers of all proficiency levels. So, further studies are needed in this regard before the results could be safely generalized.

Concerning Task 2 which indicated high variation in the scores obtained on different versions, a number of explanations may be offered. First of all, this finding could probably be justified when the nature of the two tasks is considered. Task 1 seems to be of more fixed a nature than Task 2. It does not lend itself to much flexibility and variation in writing. Therefore, if someone receives adequate practice with this task, his performance would not probably include much variation across different versions of the same task. That was the case in the present study. The first time the instructor asked the participants to focus on this task, it was difficult for many of them and they stated that they did not know how to start or end their description. They were confused as to what points to focus on. For instance, they asked the instructor whether it was better to describe all the numbers in a table or just focus on the ones that seemed more important or still more to look for certain patterns. However, after receiving instruction and training, they found the task easier to do. They stated that all the tables were very similar and more or less the same thing should be mentioned about each table. They believed that knowing certain strategies about this task would help a lot as the language which is needed to be produced for different versions of this task is the same.

Another explanation for the high similarity of performance on Task 1 may come from the limited vocabulary and structures that are usually required for this task. Knowing certain words or phrases that apply to many different examples of Task 1 could greatly improve one's performance on this task. This is also true about the grammatical structures needed. For example, present tense is the mostly-needed verb tense to complete this task, so it is much easier for students to have practice, and they are sure that this practice will prove influential as there would be a strong relationship between what they practice and what they are given in the testing situation. This could also be traced back to the impressions the participants of the present study had about the two tasks. They found Task 1 easier than Task 2, after receiving the instruction. They also found both versions of Task 1 to be of the same difficulty and believed that very similar descriptive phrases and sentences could be used in both versions to describe a graph.

In contrast to Task 1 (integrated task), which seems to be of lower score variation across different versions and for which preparation seems highly

effective, Task 2 (independent task) is more dynamic and does not easily lend itself to preparation. This does not mean that we have no idea of what we are subjected to in Task 2; rather, it means that this task is of higher flexibility and variation and calls for more creativity. Although language learners could receive instruction concerning how to write an argumentative essay and could be taught strategies to succeed in such a task, the task calls for more creativity on their part at the time of writing depending on the prompt given. That is, the arguments provided for or against something and accordingly the language produced may differ depending on the prompt. This variation is lower in Task 1.

Furthermore, the range of vocabulary and structures needed in an argumentative essay may basically depend on the requirements of the task and the prompt given. During the term, most of the participants found Task 2 more demanding and stated that their preparation was not as effective as it was for Task 1. They mentioned that they were familiar with the procedures of writing an argumentative essay; that for example, they should state the thesis statement in the introduction or that they should provide logical arguments and try to deal with the counter-arguments, etc. However, they stated that they needed more creativity and also higher knowledge of vocabulary and structure to be able to write a good essay at the time of exam. Whereas in Task 1, more vocabulary and structures tended to repeat themselves, Task 2 asked for more diversity.

Still another explanation for the difference in performance on different versions of Task 2 may come from interest in the topic or prompt. Whenever the instructor provided the participants with a topic, he asked their idea about the topic. It was not unexpected that they had diverse interests and the instructor found it very difficult to find a topic which was more or less equally interesting to all the participants. This variation in interest was less of a problem in Task 1 because different prompts have substantial similarities due to the nature and purpose of this task.

Finally, topic familiarity would provide a good justification for the results of the present study. It is quite clear that one can write better on a familiar topic. In a study of graph-writing task, Yang (2012) found that graph familiarity, topical knowledge, and test-wiseness strategy use were the possible sources of construct-irrelevant variance. Of course, IELTS administrators state that the topics are interesting, appropriate for, and easily understood by candidates; and that they are continuously pretested to ensure comparability and equality (IELTS, 2002). They argue that they benefit

from “both expert judgments by academic staff from the target domain and empirical approaches to match the test tasks with the target domain tasks and to achieve high construct representativeness and relevance” (Uysal, 2010, p.317). However, it seems that such attempts are not efficient enough. For instance, Mickan, Slater, and Gibson (2000) found that the lexicogrammatical structures in the prompts had an impact on the task comprehension and writing performance. In fact, controlling for the effect of topic/prompt is very difficult as test takers may come from many diverse backgrounds and reading experiences of the topic (Kroll & Reid 1994). The participants in the present study had a better performance on the second topic (advantages and disadvantages of the Internet). It is possible that they found this topic more interesting or that they had more information about this topic. This seems logical, as today almost everybody, especially the young generation, is expected to be familiar with the Internet and finds something interesting on it. Therefore, one can assume that the participants of the present study were probably more interested in and familiar with the topic of the Internet than the first topic which asked them to elaborate on the idea of spending money to go to other planets. Thus, better performance on the former topic is not surprising. The literature on L2 writing indicates that topic familiarity and interest can affect one’s writing performance (e.g., Cho, 2003; Cho et al., 2013; Gebril, 2009; He & Shi, 2012; Jennings, Fox, Graves, & Shohamy, 1999; Kroll & Reid, 1994; Schoonen, 2005). This means that test takers’ performance on Task 2 may differ significantly depending on topics/prompts, and hence the scores are not necessarily comparable and generalizable. This can reduce the validity of the score interpretations and decisions because they are based on scores which are not truly representative of the test takers’ abilities. This becomes more problematic when we take into account the fact that more weight is given to Task 2 in assessing the writing ability of IELTS candidates.

5. Conclusions and Implications

The present study focused on IELTS academic writing tasks to see how prompt and task type can affect test takers' performance. As the Writing Task 1 is a multifaceted task in terms of requiring graphical literacy in addition to writing ability and can be well considered as an integrated task that resembles academic genre of writing, the assumption was that it would lead to a different writing ability estimation than that of Writing Task 2, which is a traditional independent essay writing task. The results, however,

revealed that the two tasks may elicit similar performance on the part of the test takers when the overall writing quality is of concern and the same rubric is used in scoring the essays. The scores obtained on the two tasks were not significantly different.

Furthermore, a significant variation in scores was found in Task 2 because of the prompt effect. Topic familiarity and interest could provide an explanation for this variation. In contrast, lack of a significant variation in scores on Task 1 may be due to the less flexible nature of this task, as compared with Task 2. This can provide evidence for the higher consistency of scores obtained from integrated tasks than those of independent tasks, as unlike Task 2, Task 1 indicated a very high level of consistency in the scores obtained from the two versions of this task. Previously, Yang (2012) had also found validity evidence for graph-writing test tasks.

It can also be concluded from the results of this study that instruction is really effective in reducing the error variance that may come from the construct-irrelevant dimensions of integrated tasks. Although graph description task may create bias against those test takers who lack graphical schema, and therefore, their writing ability may be underestimated because of this, it seems that due instruction and practice with samples of such tasks can prove highly effective in removing this bias and creating more test fairness as all the test takers, regardless of their background knowledge, start their writing from the same source. Before the instruction, many of the participants of the present study had problems understanding the tables and graphs in Task 1; however, after receiving instruction enriched with feedback, many of them were pleased with the task and found it more or less easy to do. It is, however, worth mentioning that even after the instruction, not all the problems related to graph description were solved. For example, in the consistent feedback sessions held after each writing task, some students stated that they confused percentage with raw numbers or that in tables with several patterns to discuss, they were not sure which patterns were more important to focus on in the limited time given. But, overall, it seems that instruction in the present study helped remove the problem of multi-faceted-ness of the integrated tasks (Charge & Taylor, 1997; Fox, 2003; Upshur & Turner, 1999) by providing the participants with the graphical schemata needed. As such, unlike the assumption behind the study, it was Task 2 (independent task) that even after instruction suffered from inconsistency of the results, because of the impact of construct-irrelevant factors such as topic familiarity and interest. The study indicated

that the effect of such factors cannot be easily removed even through instruction, because controlling what background knowledge and interests test takers bring to the testing situation is a highly challenging task.

The findings of this study in line with the studies such as Friel, Bright, & Curcio (2001) and Yang (2012) call for more attention on the part of teachers to be given to students in teaching them the graphical literacy needed to be able to show their best of writing. The students may explicitly be taught the strategies that are helpful in comprehending and describing graphs and tables.

Also, test developers are invited to consider the effect of time in graph-description tasks. Although time could be a factor in performance on independent tasks as well, it seems to be a more problematic issue in performance on integrated tasks such as graph description tasks because the test takers have only 20 min to both interpret the graph and to write, but in the independent task they have 40 min to write about a topic. In the present study, although students had a good performance on the integrated task of graph description, still there were some students who suffered from lack of time and stated that they could not adequately deal with the graph description in such a short time (20min) as the graph analysis sometimes took much of their time and did not leave them with enough time for writing. Shah and Freedman (2009) consider the solution to be providing more time in such writing tasks as the inexperienced graph viewers need more time to process the information in a graph.

Finally, it should be noted that although independent writing tasks were found to suffer from inconsistency of the results, this does not imply that we should omit such tasks from standardized high-stakes tests such as IELTS or even any other writing tests. Given that independent writing tasks by nature are integral to academic writing, and students in different disciplines are oftentimes required to write argumentative essays, elimination of such writing tasks from high-stakes tests is not practical. Instead, the findings of studies such as the current one through shedding light on the intervening factors in writing assessment help us to continuously revise and restructure such tests so that they would meet the requirements of sound and scientific testing practices. Furthermore, given that such intervening factors are always at work, caution needs to be applied to the interpretations made of scores on such tests.

6. Limitations of the Study

Like any other study, the current study needs caution in generalizing the findings due to the limitations of the study. First, the writing tasks used in this study were taken from mock tests of IELTS. Using original IELTS tests may bring about different results. Second, given the variation of formats in Task 1, the results are limited to the formats used in the study. It is likely that using other formats will produce different results. Finally, the study may suffer from the sampling procedure adopted in the study because convenience sampling was employed.

References

- Brown, J. D. H. (1998). *An investigation into approaches to IELTS preparation, with particular focus on the Academic Writing component of the test* (IELTS research reports Vol. 1, pp. 20-37). Canberra: IELTS Australia.
- Charge, N., & Taylor, L. B. (1997). Recent developments in IELTS. *ELT Journal*, 51(4), 374-380.
- Cho, Y. (2003). Assessing writing: Are we bound by only one method? *Assessing Writing*, 8(3), 165-191.
- Cho, Y., Rijmen, F., & Novak, J. (2013). Investigating the effects of prompt characteristics on the comparability of TOEFL iBT™ integrated writing tasks. *Language Testing*, 30(4), 513-534.
- Cohen, J. W. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Elder, C., & O'Loughlin, K. (2003). *Investigating the relationship between intensive EAP training and band score gain on IELTS* (IELTS research reports Vol. 4, pp. 207-254). Canberra: IELTS Australia.
- Ellis, M., Chong, S., & Choy, Z. (2013). IELTS as an indicator of written proficiency levels: A study of student teachers at the National Institute of Education, Singapore. *International Journal of Educational Research*, 60, 11-18.
- Fox, J. (2003). From products to process: An ecological approach to bias detection. *International Journal of Testing*, 3(1), 21-48.
- Friel, S., Bright, F., & Curcio, G. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education*, 32(2), 124-158.

- Gao, X., & Brennan, R. (2001). Variability of estimated variance components and related statistics in a performance assessment. *Applied Measurement in Education, 14*(2), 191-203.
- Gebril, A. (2006). *Independent and Integrated academic writing tasks: A study in generalizability and test method* (Unpublished doctoral dissertation). The University of Iowa, Iowa City, USA.
- Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing, 26*(4), 507-531.
- Gebril, A., & Plakans, L. (2009). Investigating source use, discourse features, and process in integrated writing tests. *Spain Fellow Working Papers in Second or Foreign Language Assessment, 7*, 47-84.
- Golder, K., Reeder, K., & Fleming, S. (2011). Determination of appropriate IELTS writing and speaking band scores for admission into two programs at a Canadian post-secondary Polytechnic institution. *The Canadian Journal of Applied Linguistics, 14*(1), 222-250.
- Green, A. (2005). EAP study recommendations and score gains on the IELTS Academic Writing test. *Assessing Writing, 10*(1), 44-60.
- Green, A. (2007). Washback to learning outcomes: A comparative study of IELTS preparation and university pre-sessional language courses. *Assessment in Education: Principles, Policy & Practice, 14*(1), 75-97.
- Grabe, W. (2001). Reading-writing relations: Theoretical perspectives and instructional practices. In D. Belcher & A. Hirvela (Eds.). *Linking literacies: Perspectives on L2 reading-writing connections* (pp. 15-47). Ann Arbor, MI: University of Michigan Press.
- Guo, L. (2011). *Product and process in TOEFL iBT independent and integrated writing tasks: A validation study* (Unpublished doctoral dissertation). Georgia State University, Georgia, USA.
- Hamp-Lyons, L. (1991). Pre-text: Task-related influences on the writer. In L. Hamp-Lyons (Ed.). *Assessing second language writing in academic contexts* (pp. 87-110). Norwood, NJ: Ablex.
- Hamp-Lyons, L., & Kroll, B. (1996). Issues in ESL writing assessment: An overview. *College ESL, 6*(1), 52-72.
- He, L., & Shi, L. (2012). Topical knowledge and ESL writing. *Language Testing, 29*(3), 443-464.
- International English Language Testing System [IELTS]. (2002). *The IELTS handbook*. Cambridge: University of Cambridge Local Examinations Syndicate, the British Council, IDP Australia.

- Jennings, M., Fox, J., Graves, B., & Shohamy, E. (1999). The test-takers' choice: An investigation of the effect of topic on language-test performance. *Language Testing*, 16(4), 426-456.
- Kroll, B., & Reid, J. (1994). Guidelines for designing writing prompts: Clarifications, caveats, and cautions. *Journal of Second Language Writing*, 3(3), 231-255.
- Lee, Y., & Kantor, R. (2005). *Dependability of new ESL writing test scores: Tasks and alternative rating schemes* (TOEFL Monograph Series no. 1). Princeton, NJ: ETS.
- Mackey, A., & Gass, M. G. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- McCarter, S. (2002). *Academic writing practice for IELTS*. Retrieved from <http://www.scribd.com/doc/70186089/Academic-Writing-Practice-for-Ielts-2002-Sam-Mccarter>.
- Mickan, P., Slater, S., & Gibson, C. (2000). Study of response validity of the IELTS writing subtest. In R. Tulloh (Ed.). *International English language testing system* (vol. 3, pp. 29-48). Canberra, Australia: IELTS Australia.
- Moore, T., & Morton, J. (1999). *Authenticity in the IELTS academic module writing test: A comparative study of Task 2 items and university assignments* (IELTS Research Reports no. 2.). Canberra: IELTS Australia.
- Moore, T. & Morton, J. (2005). Dimensions of difference: A comparison of university writing and IELTS writing. *Journal of English for Academic Purposes*, 4(1), 43-66.
- Moore, T. & Morton, J. (2007). Authenticity in the IELTS academic module writing test: A comparative study of Task 2 items and university assignments. *Studies in Language Testing*, 19, 197-248.
- O'Loughlin, K., & Wigglesworth, G. (2003). *Task design in IELTS academic writing Task 1: The effect of quantity and manner of presentation of information on candidate writing* (IELTS Research Reports, Vol. 4, Paper 3). Canberra, Australia: IDP IELTS Australia.
- Plakans, L. (2007). *Second language writing and reading-to-write assessment tasks: A process study* (Unpublished doctoral dissertation). The University of Iowa, Iowa City: IA.
- Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing*, 26(4), 561-587.

- Rao, C., McPherson, K., Chand, R., & Khan, V. (2003). *Assessing the impact of IELTS preparation programs on candidates' performance on the GT reading and writing test modules* (IELTS research reports Vol. 5, pp. 237-262). Canberra: IELTS Australia.
- Read, J., & Hayes, B. (2003). *The impact of the IELTS test on preparation for academic study in New Zealand* (IELTS research reports Vol. 4, pp. 153-206). Canberra: IELTS Australia.
- Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, 22(1), 1-30.
- Shah, P., & Freedman, E. G. (2011). Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in Cognitive Science*, 3(3), 560-578.
- Spack, R. (1993). Student meets text, text meets student: Finding a way into academic discourse. In J. Carson & I. Leki (Eds.). *Reading in the composition classroom: Second language perspectives* (pp. 183-196). Boston, MA: Heinle & Heinle.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson Education.
- Upshur, J. A., & Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing*, 16(1), 82-111.
- Uysal, H. H. (2010). A critical review of the IELTS writing test. *ELT Journal*, 64(3), 314-320.
- Weigle, S. (2002). *Assessing writing*. Cambridge: Cambridge University Press.
- Weigle, S. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing writing*, 9(1), 27-55.
- Yang, H. C. (2012). Modeling the relationships between test-taking strategies and test performance on a graph-writing task: Implications for EAP. *English for Specific Purposes*, 31(3), 174-187.