

TEST METHOD EFFECT ON WRITING PERFORMANCE

Liu Yang

National Educational Examinations Authority , Ministry of Education

Abstract

The main objective of this paper was to investigate the test method effect of two writing tests on the performances of sixty sophomore English majors from China University of Mining and Technology , who were divided randomly into two experiment groups and took a reading integrated writing test and a timed impromptu essay test respectively . Their essays were rated anonymously by two independent raters using the same rating scale with categories that measured content , organization , accuracy and vocabulary . Besides examination of the reliability , Many Faceted Rasch analysis was applied to probe into the influence of domain difficulty and rater severity . Results of the comparison revealed four important findings . First , both two tests were reliable to be fair measures for assessing writing abilities . Second , significant difference was found between two groups in content , organization , and vocabulary , but no difference was observed in accuracy . Third , the reading tasks facilitated participants in generating ideas , organizing essays and using vocabularies . Finally , compared with the participants in the reading integrated writing test , participants in the timed impromptu essay test met difficulties in using their vocabularies in the writing process . Findings implicated that 1) a reading integrated writing test could be an alternative to a timed impromptu essay test in academic contexts , and 2) much more investigation was still needed to probe into the writing process and read to -write process .

Key words

writing assessment ; read to -write ; method effect

1 . Backgrounds

Historically , writing assessment has changed in three overlapping waves in the past fifty years (Yancey 1999) . In the first wave (1950 -1970) emphasis was laid on reliability of the measurement in a form of indirect objective tests ; in the second wave (1970 -1986) validity became dominant , and writing assessment took a form of holistically scored essay ; and in the current wave (1986 -present) experts paid attention to both reliability and validity and portfolio assessment and programmatic assessment were used in assessment . Therefore , the solution to the existing problems in writing assessment is always to seek the balance among different interested groups (White 1994) .

Many studies (Crusan 2002 ; White 1994 ; Yancey 1999) have criticized the indirect objective writing test for its disjunction between classroom practice and testing practice . Some other studies (Alderson & Hamp -Lyons 1996 ; Crusan 2002 ; Cumming 2002 ; Hamp -Lyons 1998) also point out the weakness of " one shot " timed impromptu essay tests for it " may cause students to misdirect their learning in preparation for the test to focus on this one type of composition , rather than other types of writing that might be more important for their studies ; or students can be easily coached into producing a simple form of this composition without truly developing other aspects of their writing abilities " (Cumming 2002 : 77) . Furthermore , though portfolio assessment is regarded as a more reliable and accurate approach to assessing writing ability , it also restrains itself from being popular due to its low efficiency and low

feasibility in administration .

Based on those facts ,the reading integrated writing test was proposed as an alternative to the timed impromptu essay test in academic contexts for its advantages of authenticity over the timed impromptu essay test and of economy over the portfolio writing test . Weigle (2004) has pointed out several advantages of this read to -write test . It has been proved that academic writing is rarely done in isolation but with source texts ,and these source texts bring test takers on a more equal footing by providing them with a common information source which activates writers knowledge around a topic and helps them generate ideas for their writing .

This study investigates the methods used to measure writing ability of intermediate EFL learners in academic contexts ,and it attempts to probe into the effects of different testing methods on assessing writing ability by comparing a timed impromptu essay test with an alternative reading integrated writing test . In this study ,the aim is to work out an alternative assessing method of writing ability to the timed impromptu essay test in academic contexts in China . Hopefully ,the findings of the study would shed lights on the understanding of writing process and provide insights and have some implications in Chinese EFL teaching ,especially academic writing instruction .

Our research questions are :

- 1) Do participants perform differently between a reading integrated writing test and a timed impromptu essay test in both overall performance and four domains —content ,organization , accuracy and vocabulary —of writing tests ?
- 2) If they do ,do the reading tasks hinder or facilitate the participants in their writing performances ? If they do not ,what are the possible reasons ?

Then the research questions generate some predictions concerning both two test methods .

Hypothesis 1 :

Participants in the reading integrated writing test perform better than those in the timed impromptu essay test .

Hypothesis 2 :

The integrated reading tasks facilitate participants performances in the test .

2 . Research Design

In this paper ,test method was the between group factor which had to be taken into account in the study as the independent variable . Different test methods in this study were defined as the different tasks test takers needed to accomplish before they finished their essays on the same topic . The dependent variable was the score that test takers obtained in every component of the test as well as a total score added up from those components .

The research design was of a 1 x2 design with test method as the between group factor . All the participants were randomly divided into two groups by the different tests they were going to take . One group took a reading integrated writing test ,in which every participant was presented with two passages reflecting opposite perspectives of the same issue ,and the other group took a timed impromptu essay test with a bare prompt (no reading tasks before writing) . Participants in the first group were required to read the passages before writing ,and another 45 minutes were given to finish reading tasks .

After tests ,the finished essays were rated by two independent raters and the scores they gained in different components together with the total scores were for statistical analysis .

Table 2-1 . Research design

Testing Method	
Reading Integrated Writing test	Group 1
Timed Impromptu Essay Test	Group 2

2.1 Participants

Participants were solicited from sophomore English majors of China University of Mining and Technology. They were all Non-Native Speakers (NNS) of English taking English as their major and speaking Chinese as their mother tongue. The average age of these participants was 19. These participants had finished senior middle school and received two years of college English instruction who could be treated as intermediate English learners and potentially would conduct academic research in the future. As sophomore English majors, they were required to obtain the basic writing skills and could write 100-300 words argumentative essays in English. Besides that, they all should take TEM 4 (Test for English Majors Band 4) proficiency test of English in the second year. In this study, all the participants had passed the TEM 4 proficiency tests.

The total 60 participants, each with an assigned number to indicate his/her identification, were divided into two groups randomly and tested by two tests respectively. As typical in China, the majority of these English majors were female. 30 participants were in each group, with 18 females and 12 males in Group 1 taking reading integrated writing test; and with 23 females and 7 males in Group 2 taking timed impromptu essay test.

Each group took one test only. After the test, all the papers were collected and transmitted to raters and rated anonymously by using the same scoring rubric.

2.2 Raters

Both two raters named Liu and Luan had experience of rating essays in PETS (Public English Test System) and NMET 2 (National Matriculation English Test 2) writing tests. Liu was the author of this paper, too.

In this study, both two raters were trained for several hours to be familiar with the rating scale applied to both writing tests. The training process was mainly focused on the rating scale by discussion. Before rating both two raters studied and discussed the rating scale, and then ten essays (five from each test) were selected and mixed together for rater training. No more than one point of difference in each domain was required at the end of the rater training process. When disagreement appeared, two raters discussed about the difference and necessary adjustment was made. Due to the size of the sample, a larger sample would be helpful to reach an ideal result of rater training.

2.3 Instruments

Since all the 60 participants were divided into two groups, no significant difference in language proficiency between two groups was expected before experiment. To achieve this goal, the TEM 4 scores were used to investigate the language proficiency difference before the experiment.

The instruments of this study to measure participants' writing ability were two writing test papers. The first one — Writing Test Paper A (see Appendix A) with reading tasks, was adapted directly from Weigle (2004) by removing the questions following the reading tasks. Removing those questions was for two reasons: the time for this experiment was limited, and the author believed that those questions might distract participants' attention from writing to answering questions.

Writing Test Paper A was made up of two parts: Part 1 and Part 2. In Part 1, two reading passages on the same topic were presented. Both two passages were about 300 words long and about the same topic but opposite perspectives were held in the two passages: one argued that genetically modified plants were dangerous to our health and environment; the other believed that genetic engineering was an important tool in feeding the world's population. After reading, participants were required to write an essay about 300 words by supporting one of the positions in Part 2. Besides, a clear demand of no copy from the source texts was added in the instructions of the test paper.

The other paper — Writing Test Paper B (see Appendix B) used in this study was adapted by the author himself from Writing Test Paper A. In this test, reading tasks were eliminated, but the same topic was kept. Because there were no reading tasks, slight changes of wording in directions and requirements

were made . Words referring source texts were erased ,but directions about using writers own ideas to support the argumentation were added like most of the timed impromptu essay tests .

All of the essays produced in both two tests were rated by the same two raters using the same rating scale (see Appendix C) . The rating scale was also adapted from Weigle (2004) ,in which some points about copying reading materials in writing essays were deleted because of its application into the rating process of the test paper Bin this study . It was an analytical rating scale with four different domains : content ,organization ,accuracy ,and vocabulary range ,and it had ten points in 5 ranks for each domain .

2.4 Procedures

To make sure the homogeneity in language proficiency of the two groups before the experiment ,all the participants TEM 4 scores were computed by SPSS13 .0 using an independent sample t test analysis .

After that ,the two groups were tested in two different classrooms at the same time of two different days in the same week . At the beginning of the tests all the participants were clarified with the directions and requirements of tests .

During the tests ,Group 1 was given an extra 45 minutes to finish the reading tasks ,and Group 2 was required to finish their tests in 55 minutes . Both groups were required to write their essays in answer sheets . If participants in Group 1 finished the reading tasks in less than 45 minutes ,they could start the writing task immediately ,but this might give participants in Group 1 advantage over the participants in Group 2 in the testing time to finish the writing task .

The collected essays were copied for two raters . For each rater ,all the essays from two tests were mixed together without name ,and the same rating scale was used to rate all the essays from the two tests . Two raters took one day to rate all the essays independently at the same time . After rating ,all the essays were rearranged to two groups and were examined by two raters . When radical ratings with a gap of more than two points in the scores of each domain appeared between two raters ,two raters discussed about the rating score based on the rating scale ,and adjustments were made .

2.5 Analytical Method

Besides the independent sample t test analysis conducted before the experiment ,cronbach alpha and correlation coefficients were computed to determine the internal consistency and rater consistency between two raters among all the subcomponents in two tests .

To answer those research questions above ,a two -way ANOVA analysis was calculated to indicate the differences of the participants performances in two groups ,and the following Many -Faceted Rasch analysis presented more information about the two tests .

SPSS version 13 .0 was used for computing independent sample t test ,the internal consistency reliability for the two raters ,correlations ,and two -way ANOVA . The Many -Faceted Rasch analysis was conducted using the computer program Facets ,version 3 .58 for PC .

3 . Results and Discussion

Before the experiment ,it is necessary to make sure that the two different groups are not different in their language proficiency in English . The result of t test indicates no significant difference between the two groups in language proficiency ($t(58) = 1.399$, $p = .167$) ,though both Mean score and Std . Deviation are different .

Table 3 1 . Results of t test for two groups in TEM 4 tests

Group	n	Mean	SD	t	df	Sig (2 tailed)	Mean Difference
1	30	69 .1667	6 .64926	1 .399	58	.167	2 .56667
2	30	66 .6000	7 .53658				

Since no difference of language proficiency occurred between two groups ,it is also necessary to examine the rater reliability and the reliability of two tests first . The rater reliability can be examined by the intra rater reliability and inter rater reliability . Intra rater reliability is the consistency within each rater s ratings ,and inter rater reliability is the consistency across each rater s ratings (Bachman 1990) . The intra rater reliability can be examined by the infit mean square from the many faceted analysis of raters . The infit mean square indicate raters internal self consistency . McNamara (1996) suggests that an acceptable range is the mean \pm twice the standard deviation . The infit mean square in the last column of Table 3 2 and Table 3 3 indicates no rater was identified as misfitting ; fit values are within the range of two standard deviations around the mean ((0 .94 \pm [0 .04 \times 2] for Group 1 ,(0 .86 \pm [0 .08 \times 2] for Group 2) . In other words ,both two raters are self consistent in scoring ,and intra rater reliability is achieved .

Table 3 2 . Calibration of rater facet for Group 1

N	Rater	Rater Severity Measure (in logits)	Standard Error	Infit Mean Square Index
1	A Rater	-.08	.14	.97
2	B Rater	-.50	.14	.90
	Mean	.29	.14	.94
	SD	.04	.00	.04

Reliability of separation index = .56 ;
 Fixed (all same) chi square :4 .6 df :1 ,significance :p = .03

Table 3 3 . Calibration of rater facet for Group 2

N	Rater	Rater Severity Measure (in logits)	Standard Error	Infit Mean Square Index
1	A Rater	.45	.15	.94
2	B Rater	-.02	.15	.79
	Mean	.21	.15	.86
	SD	.24	.00	.08

Reliability of separation index = .59 ;
 Fixed (all same) chi square :4 .9 df :1 ,significance :p = .03

Besides the intra rater reliability ,inter rater reliability is reported in Table 3 4 and confirmed by ANOVA analysis . From Table 3 4 ,strong positive correlations (p < .01)are found between two raters measuring the same component (content :r (58) = .857 ,p < .01 ;organization :r (58) = .836 ,p < .01 ;accuracy :r (58) = .850 ,p < .01 ;and vocabulary :r (58) = .851 ,p < .01)indicating that two rater are consistent in the rating process . Significant correlations (p < .01) were also observed among different components measured by the same rater ,suggesting that four traits in this study measure different aspects of the same writing ability . Besides ,results of ANOVA analysis shown in Table 3 5 and Table 3 6 present no difference of raw scores between two raters . They all prove the ratings between two raters are consistent .

Table 3 4 . Correlations of the ratings for two raters

		Rater A				Rater B			
		Co	Or	Ac	Vo	Co	Or	Ac	Vo
Rater A	Co	1 .00							
	Or	.864 **	1 .00						
	Ac	.781 **	.832 **	1 .00					
	Vo	.798 **	.827 **	.780 **	1 .00				

		Rater A				Rater B			
		Co	Or	Ac	Vo	Co	Or	Ac	Vo
Rater B	Co	.857 **	.838 **	.828 **	.835 **	1.00			
	Or	.809 **	.836 **	.840 **	.810 **	.845 **	1.00		
	Ac	.865 **	.870 **	.850 **	.797 **	.850 **	.811 **	1.00	
	Vo	.797 **	.835 **	.801 **	.851 **	.809 **	.748 **	.892 **	1.00

** Correlation is significant at the 0.01 level (2 tailed).

Co = Content ;Or = Organization ;Ac = Accuracy ;Vo = Vocabulary

It should be noted that the second column of Table 3-2 and Table 3-3 shows severity measures of two raters in logits. The severity span between two raters in Table 3-2 is 0.42 logit and 0.47 logit in Table 3-3. The reliability of separation index indicates the likelihood to which raters consistency differs from one another in overall severity, and a low reliability is desired for raters, because ideally different raters would be equally severe (Park 2004 ;Liu 2005). In this study, the reliability indices are 0.56 for Group 1 and 0.59 for Group 2. Also, the chi square of 4.6 (df = 1) for Group 1 and 4.9 (df = 1) for Group 2 are significant at p < .05, which indicate difference of rater severity. Contrary to this, the results of two-way ANOVA analysis in Table 3-6 show no statistical differences between two raters. Unlike ANOVA analysis Many-Faceted Rasch model provides more details about the difference of rater severity by comparing the observed score with the expected score. Thus, from Many-Faceted Rasch analysis we obtain more information about rater severity difference than ANOVA analysis which compares the raw scores of the two raters. Because one of the two raters is also the designer and the author of this paper, the rater severity difference is caused by researcher effect to some degree, and another independent rater would be helpful to improve the rating. Although with rater severity differences between two raters, the intra rater and inter rater reliability are still achieved.

Finally, internal consistency reliability is one important concern for any test, in that it indicates the reliability of discrimination for each test. Hence, the cronbach alpha reliability is computed to show the internal consistency reliability. And the higher the reliability is, the more reliable the test discriminates participants abilities. The cronbach alpha reliability for Group 1 taking reading integrated writing test is 0.934, and that for Group 2 taking timed impromptu essay test is 0.951, both of which prove that the two writing tests are reliable to discriminate participant writing abilities.

Table 3-5. Descriptive statistics across two groups in four domains

	Group	n	Mean	SD	Mini	Max
Total	1	60	24.9667	5.65376	12.00	34.00
	2	60	21.9500	5.72069	10.00	32.00
Co	1	60	6.0500	1.44298	3.00	9.00
	2	60	5.1167	1.46243	2.00	8.00
Or	1	60	6.0333	1.52900	3.00	9.00
	2	60	5.3333	1.63299	2.00	9.00
Ac	1	60	6.4667	1.65157	3.00	9.00
	2	60	5.9000	1.53711	2.00	9.00
Vo	1	60	6.4167	1.55456	3.00	9.00
	2	60	5.6000	1.48666	3.00	9.00

Co = Content ;Or = Organization ;Ac = Accuracy ;Vo = Vocabulary

Table 3-6 . Two-way ANOVA for different test methods

Source	df	MS	F	Sig .
Total				
Between group :Method	1	273 .008	8 .331	.005 * *
Within group :Rater	1	15 .408	.470	.494
Method * Rater	1	.008	.000	.987
Content				
Between group :Method	1	26 .133	12 .247	.001 * *
Within group :Rater	1	.300	.141	.708
Method * Rater	1	1 .200	.562	.455
Organization				
Between group :Method	1	14 .700	5 .787	.018 *
Within group :Rater	1	.300	.118	.732
Method * Rater	1	.300	.118	.732
Accuracy				
Between group :Method	1	9 .633	3 .792	.054
Within group :Rater	1	4 .033	1 .588	.210
Method * Rater	1	1 .633	.643	.424
Vocabulary				
Between group :Method	1	20 .008	8 .625	.004 * *
Within group :Rater	1	3 .675	1 .584	.211
Method * Rater	1	.208	.090	.765

* * $p < .01$; * $p < .05$

In above discussion the reliability of raters and tests has been examined ,and then the discussion will focus on the two research questions proposed at the beginning . After the experiment ,a quantitative difference is expected to occur in the employments of different test methods in writing assessment . The results of two-way ANOVA in Table 3-6 show that significant statistical differences between two groups appear in both overall performances and three of the four domains :content , organization , and vocabulary ,and no significant difference is observed in the domain of accuracy . Besides ,we also note a different significant level between in organization ($p < .05$) and in content ($p < .01$) together with vocabulary ($p < .01$) . These results prove that the different test methods affect participants greatly on their performances in content ,organization ,and vocabulary .

The difference in content is expected .it is assumed that the reading tasks can relief the cognitive burden for interpreting the source texts materials ,then save the working memory capacity for some attention consuming ,higher level processes of writing like planning the organization of the essay (Hayes

1996). Ideally ,difference between groups appears in content . Group 1 taking reading integrated writing tests got higher means than Group 2 ,which confirms the expectation that the reading tasks provide participants with ideas about the contents and even may trigger participants knowledge for more ideas . To some degree ,it eases students anxiety of finding contents for writing . Consequently ,essays from the Group 2 use Mr . Yuan Long Ping (the father of cross bred rice in China) more frequently as examples to illustrate their ideas compared with Group 1 . Differences between these two writing processes are worth investigating ,but limited by the resources of test administration ,we did not use any think aloud process or interview to provide more information about it .

Although the difference is also found in organization ,it is not as significant as those in content and vocabulary . Unlike speaking ,writing ability is affected greatly by the education processes (Campbell 1999 ; Weigle 2002) . Since writing instructions in China treat writing ability as a discrete skill and focus mainly on the test taking skills (You 2004) . It is not unusual to find out that students in the same educational system may not vary too much from each other ,even though English majors may put more effort into language practicing and training of overall language competence . Written essays from both tests support this assumption ,too . Most of the passages were written in the almost the same pattern with four paragraph and similar introduction . These reasons explain why the different level in organization is not as the same as the other two domains .

It is surprising to find the difference between two groups in vocabulary . It is assumed that the integrated reading tasks affect participants on generating ideas and organizing essays ,not on the linguistic knowledge like vocabulary ,so no difference in vocabulary is expected . One explanation for this may be that the reading tasks provide participants with more various words about the topic which can be retrieved in writing process . Similarly ,another explanation is that reading materials trigger participants vocabulary knowledge and activate it like triggering ideas in the writing process . To prove these assumptions ,more efforts are needed to investigate both test taking and writing processes .

The only domain which shows no difference between two groups is accuracy . As English majors ,participants in this study have put more effort in language practicing and training of overall language competence than those non -English majors . After two years of language training ,they should not have many problems with the gram mar and can generate gram matical sentences in their essays . In addition ,the reading tasks integrated in the writing tasks are not aimed to affect participant performances in gram mar . Therefore ,no difference should appear in accuracy ,if there is no difference in language proficiency between two groups . ANOVA analysis confirms the expectation by showing no difference in accuracy . From Many Faceted Rasch analysis ,the fact that participants in two tests all find the accuracy is easiest domain to accomplish also partially supports it (-.42 in logits for Group 1 and -.88 in logits for Group 2) .

In order to answer the second research question ,we need to take means of each group into account . ANOVA analysis shows that Group 1 taking the reading integrated writing test obtained higher mean scores in all the domains ,and the differences between two groups are statistically significant in content ,organization and vocabulary . Since Group 1 has already gained higher mean scores than Group 2 in those domains ,it is safe to draw a conclusion that the reading tasks have positive effects on those domains ,and they can facilitate participants to form ideas and use more various words in writing process ,but its effects on organization may be limited .

Table 3 -7 . Calibration of domain facet for Group 1

Domain	Difficulty Measure (in logits)	Standard Error	Infit Mean Square Index
Content	.42	.20	.96
Organization	.23	.19	.79
Vocabulary	-.23	.20	1.02
Accuracy	-.42	.20	.98

Domain	Difficulty Measure (in logits)	Standard Error	Infit Mean Square Index
Mean	.00	.20	.94
SD	.34	.00	.09

Separation index = 1.43 ,Reliability of separation index = .67 ;
 Fixed (all same) chi square :12.0 df :3 ,significance :p = .01

Table 3-8 . Calibration of domain facet for Group 2

Domain	Difficulty Measure (in logits)	Standard Error	Infit Mean Square Index
Vocabulary	.92	.22	.97
Organization	.30	.20	.85
Content	-.34	.22	.88
Accuracy	-.88	.22	.75
Mean	.00	.21	.87
SD	.68	.01	.08

Separation index = 2.99 ,Reliability of separation index = .90 ;
 Fixed (all same) chi square :38.9 df :3 ,significance :p = .00

The different sequences of the domain difficulties are observed in second column of Table 3-7 and Table 3-8 . In Group 1 ,content is on the top of the sequence followed by organization ,vocabulary and accuracy ,but vocabulary is on the top of the sequence followed by organization ,content and accuracy in Group 2 . From the top to the bottom ,the higher its position is ,the more difficult it will be . For Group 1 ,integrating reading tasks in writing test is to make the test more authentic and help participants generate ideas in their essays . And research also indicates that good writers spend more time planning and tend to edit their writing for content and organization rather than simply making surface changes to the text (Hayes 1996 ;Hayes & Flower 1980 ;Whalen & Menard 1995) . Participants in reading integrated test find content the most difficult proves that their focus in writing is to absorb the ideas from the source texts and translate these ideas into their written language . Compared with Group 1 ,the most difficult domain is vocabulary for participants in Group 2 . It indicates that participants in Group 2 meet more difficulties in vocabulary . this finding echoes the studies demonstrating that EFL writers plan and evaluate relatively more frequently at the linguistic level compared to the textual and pragmatic level (Sasaki & Hirose 1996 ;Whalen & Menard 1995) .

4 . Conclusion

Based on the discussion above ,it can be concluded that two tests are both reliable to measure students writing ability . Participants in the two tests performed differently in the overall performances . And participants in the reading integrated writing test obtained higher scores in all the domains than participants in the timed impromptu essay test . Their performances between two tests are significantly different in content ,organization ,and vocabulary ,but no difference is observed in the domain of accuracy . It supports the idea that the reading tasks facilitate participants in generating ideas ,organizing essays and even using vocabularies ,and reading tasks help participants generate ideas for content ,and may trigger participants vocabulary knowledge to write with more various word choices . But the effect on organization is not as obvious as effects on the content and vocabulary . And finally ,reading tasks do not affect the performance in accuracy . Since a reading integrated writing test is more authentic ,it can be an alternative to the timed impromptu essay test in academic contexts .

Though this study proves that reading tasks affect participant performances positively in the writing process some limitations are needed to be presented . First ,all the participants were English majors from

the same area , the same university in the same grade with similar educational backgrounds . The homogeneity of participant makes it hard to generalize the findings of this study to the larger population , compounding with the problem is the fact that females take a large percentage of English majors , which may also keep the findings from being powerful in generalization .

Second , more raters and more rater training will be more helpful in the study . Another independent rater will improve the reliability of the tests . If a native rater is added , more helpful information will be included . And more rater training will improve the quality of the research data .

Third , it was noticed that the reading integrated writing test and the timed impromptu essay test may measure the different abilities , but these differences were not investigated in this study , nor was the equation between the two tests forms . Therefore , more efforts are needed to explore the constructs underlying these two test forms as well as the equation investigation between them .

Finally , limited by the administration sources , this study did not conduct any think aloud protocol or interview about participants test taking process , which is very important to clarify the factors in taking writing tests . More studies are desired to probe into the writing process , the read to write process , and the test taking process . The information obtained from those studies will improve the quality of this study .

References

- Alderson , J . C . & L . Hamp-Lyons . 1996 . TOEFL preparation courses : A study of washback . *TESOL Quarterly* 13 :280-297 .
- Asencion , Y . 2004 . Validation of reading to write assessment tasks performed by second language learners . Unpublished doctoral dissertation , Northern Arizona University .
- Bachman , L . F . 1990 . *Fundamental Considerations in Language Testing* . New York : Oxford University Press .
- Bachman , L . F . 2004 . *Statistical Analyses for Language Assessment* . Cambridge : Cambridge University Press .
- Bachman , L . F . & A . S . Palmer . 1996 . *Language Testing in Practice* . New York : Oxford University Press .
- Bailey , K . M . 1998 . *Learning about Language Assessment : Dilemmas , Decisions , and Directions* . Boston : Heinle & Heinle .
- Bond , T . G . & C . M . Fox . 2001 . *Applying the Rasch Model : Fundamental Measurement in the Human Sciences* . Mahwah , New Jersey : Lawrence Erlbaum Associates , Inc . , Publishers .
- Breland , H . 1996 . *Writing Skills Assessment : Problems and Prospects* . Princeton , NJ : Educational Testing Service .
- Campbell , C . 1999 . *Teaching Second Language Writing : Interacting with Text* . Thomson Learning .
- Chenoweth , N . A . & J . R . Hayes . 2001 . Fluency in writing . Generating text in L1 and L2 . *Written Communication* 18 /1 :80-98 .
- Cho , Y . 2003 . Assessing writing : Are we bound by only one method ? *Assessing Writing* 8 :165-191 .
- Crusan , D . 2002 . An assessment of ESL writing placement assessment . *Assessing Writing* 8 :17-30 .
- Cumming , A . 2002 . Assessing L2 writing : alternative constructs and ethical dilemmas . *Assessing Writing* 8 :73-83 .
- Cumming , A . , L . Grant , P . Mulcahy-Ernt & D . E . Powers . 2004 . A teacher verification study of speaking and writing prototype tasks for a new TOEFL . *Language Testing* 21 /2 :107-145 .
- Cumming , A . , R . Kantor , D . E . Powers , T . Santos & C . Taylor . 2000 . *TOEFL 2000 Writing Framework : A Working Paper (TOEFL Monograph Series Report No . 18)* . Princeton , NJ : Educational Testing Service .
- Engelhard , G . , Jr 1992 . The measurement of writing ability with a Many-Faceted Rasch model . *Applied Measurement in Education* 5 /3 :171-191 .
- Esmaili , H . 2000 . The effects of content knowledge from reading on adult ESL students written composition in an English Language Test using reading and writing modules . Unpublished doctoral dissertation , University of Toronto , Canada .
- Fulcher , G . 1999 . Assessment in English for academic purposes : Putting content validity in its place . *Applied Linguistics* 20 /2 :221-236 .
- Hale , G . , C . Taylor , B . Bridgeman , J . Carson , B . Kroll & R . Kantor . 1996 . *A Study of Writing Tasks Assigned in Academic Degree Programs (TOEFL Monograph Series Report No . 54)* . Princeton , NJ :

Educational Testing Service .

- Hamp-Lyons ,L .1991a . Basic concepts . In L . Hamp-Lyons (ed .) . Assessing Second Language Writing in Academic Contexts . Norwood ,NJ :Ablex Publishing Corporation .
- Hamp-Lyons ,L .1991b . Reconstructing " Academic Writing Proficiency " . In L . Hamp-Lyons (ed .) . Assessing Second Language Writing in Academic Contexts . Norwood ,NJ :Ablex Publishing Corporation .
- Hamp-Lyons ,L .1991c . The writers knowledge and our knowledge of the writer . In L . Hamp-Lyons (ed .) . Assessing Second Language Writing in Academic Contexts . Norwood , NJ Ablex Publishing Corporation .
- Hamp-Lyons ,L .1998 . Ethical test preparation practice :The case of TOEFL . TESOL Quarterly 32 :329 - 337 .
- Hayes ,J . R .1996 . A new framework for understanding cognition and affect in writing . In C . M . Levy & S . Ransdell (eds .) . The Science of Writing . Mahwah ,NJ :Lawrence Erlbaum Associates .
- Hayes ,J . R . & L . S . Flower .1980 . Identifying the organization of writing processes . In L . W . Gregg & E . R . Steinberg (eds .) . Cognitive Processes in Writing , pp . 31 -50 . Hillsdale , NJ :Lawrence Erlbaum Associates
- Huot ,B . A .1990 . The literature of direct writing assessment :Major concerns and prevailing trends . Review of Educational Research 60 :237 -263 .
- Jamieson ,J . ,S . Jones ,J . Kirsch ,P . Mosenthal & C . Taylor .1999 . TOEFL 2000 Framework :A Working Paper (TOEFL Monograph Series Report No .16) . Princeton ,NJ :Educational Testing Service .
- Leki ,J .1991 . A new approach to advanced ESL placement testing . Writing Program Administration 14 /3 :53 - 58 .
- Leki ,J .1992 . Understanding ESL Writers . NH :Heinemann Educational Books
- Linacre ,J . M .1989 . Many-Faceted Rasch Measurement . Chicago :MESA Press .
- Linacre ,J . M .1999a . FACETS (Version 3 .58) [Computer Program] . Chicago :MESA Press .
- Linacre ,J . M .1999b . Investigating rating scale category utility . Journal of Outcome Measurement 3 /2 :103 - 112 .
- Linacre ,J . M .2002 . Optimizing rating scale category effectiveness . Journal of Applied Measurement 3 /1 :85 - 106 .
- Lynch ,B . K . & T . McNamara .1998 . Using G theory and many-facet Rasch measurement in the development of performance assessment of the ESL speaking skills of immigrants . Language Testing 15 /2 :158 -180 .
- McNamara ,T .1996 . Measuring Second Language Performance . New York Addison Wesley Longman .
- Moss ,P . A .1994 . Validity in high stakes writing assessment :problems and possibilities . Assessing Writing 1 /1 :109 -128 .
- Myford ,C . M . ,D . B . Marr & J . M . Linacre .1996 . Reader Calibration and its Potential Role in Equating for the TWE (TOEFL Research Report No .95-40) . Princeton ,NJ :Educational Testing Service .
- Myford ,C . M . & E . W . Wolfe .2003 . Detecting and measuring rater effects using Many-Facet Rasch measurement :Part I . Journal of Applied Measurement 4 /4 :386 -422 .
- Myford ,C . M . & E . W . Wolfe .2004 . Detecting and measuring rater effects using Many-Facet Rasch measurement :Part II . Journal of Applied Measurement 5 /2 :189 -227 .
- Nystrand ,M . ,A . S . Cohen & N . M . Dowling .1993 . Addressing reliability problems in the portfolio assessment of college writing . Educational Assessment 1 /1 :53 -70 .
- Park ,T .2004 . An investigation of an ESL placement test of writing using many-facet Rasch measurement . TESOL & Applied Linguistics 4 :1 -21 .
- Pollitt ,A . & C . Hutchinson .1987 . Calibrated graded assessment :Rasch partial credit analysis of performance in writing . Language Testing 4 :72 -92 .
- Raimes ,A .1990 . The TOEFL test of written English :Causes for concern . TESOL Quarterly 24 :427 -442 .
- Read ,J .1990 . Providing relevant content in an EAP writing test . English for Specific Purposes 9 :109 -121 .
- Sasaki ,M . & K . Hirose .1996 . Explanatory variables for EFL students expository writing . Language Learning 46 /1 :137 -174 .
- Shi ,L .2001 . Native -and nonnative speaking EFL teachers evaluation of Chinese students English writing . Language Testing 18 /3 :303 -325 .
- Vaughan ,C .1991 . Holistic assessment :What goes on in the raters mind ? In L . Hamp-Lyons (ed .) . Assessing Second Language Writing in Academic Contexts . Norwood ,NJ :Ablex Publishing Corporation .

- Victori ,M .1999 . An analysis of writing knowledge in EFL composing :A case study of two effective and two less effective writers . System 27 :537 555 .
- Wang ,W . &Q . Wen .2002 . L1 use in the L2 composing process :An exploratory study of 16 Chinese EFL writers . Journal of Second Language Writing 11 :225 246 .
- Way ,D .P . ,E .G .Joiner &M .A . Seaman .2000 . Writing in the secondary foreign language classroom :The effects of prompts and tasks on novice learner of French . The Modern Language Journal 84 /2 :171 184 .
- Weigle ,S . C .1994 . Using FACETS to model rater training effects . Language Testing Research Colloquium . Retrieved January 4th ,2006 ,from http://www.eric.ed.gov/ERICDocs/data/ericdocs2/content_storage_01/0000000b/80/24/58/04.pdf
- Weigle ,S . C .2002 . Assessing Writing . Cambridge :Cambridge University Press .
- Weigle ,S . C .2004 . Integrating reading and writing in a competency test for non native speakers of English . Assessing Writing 9 :27 55 .
- Whalen ,K . &N . Menard .1995 . L1 and L2 writers strategic and linguistic knowledge :A model of multiple - level discourse processing . Language Learning 44 /3 :381 418 .
- White ,E . M .1994 . Issues and problems in writing assessment Assessing Writing 1 /1 :11 27 .
- White ,E . M .1995 . An apologia for the timed impromptu essay test . College Composition and Communication 46 :30 45 .
- Yancey ,K . B .1999 . Looking back as we look forward :Historicizing writing assessment . College Composition and Communication 50 /3 :483 503 .
- You ,X .2004 . “ The choice made from no choice ” :English writing instruction in a Chinese University . Journal of Second Language Writing 13 :97 110 .
- 刘建达 ,2005 , 话语填充测试方法的多层面 Rasch 模型分析 ,《现代外语》第2 期。

Appendices

Appendix A

Writing Test Paper (A)

Part 1 :Read the following two passages

Directions :Read the following two passages that are arguing for two different sides of the same issue . After you finish reading ,you will write an argumentative essay in part 2 . You will have 45 minutes to read .

A . 1 . A . Excerpt from an essay published by the BioDemocracy and Organic Consumers Association

Genetic engineering is a radical new technology ,one that breaks down fundamental genetic barriers —not only between species ,but between humans ,animals ,and plants . By combining the genes of dissimilar and unrelated species ,permanently altering their genetic codes ,new organisms are created that will pass the genetic changes onto their offspring . Scientists are now inserting animal and even human genes into plants or animals ,creating unimagined transgenic life forms . For the first time in history ,human beings are becoming the architects of life .

Bio engineers will be creating tens of thousands of novel organisms over the next few years . The prospect is frightening . Genetic engineering poses unprecedented ethical and social concerns ,as well as serious challenges to the environment ,human health ,animal welfare ,and the future of agriculture . The following is just a sampling of concerns .

Genetically engineered organisms that escape or are released from the laboratory could cause a great deal of harm to the environment . Genetically engineered “biological pollutants” have the potential to be even more destructive than chemical pollutants . Because they are alive ,genetically engineered products are inherently more unpredictable than chemical products —they can reproduce ,migrate ,and mutate . Once released ,it will be virtually impossible to recall genetically engineered organisms back to the laboratory .

Since scientists will never be able to ensure a 100 percent success rate ,gene splicing will likely result in unanticipated outcomes and dangerous surprises . Researchers recently found that genetically altering

plants to resist viruses can cause the viruses to mutate into new , more virulent forms , or forms that can attack other plant species . Furthermore , genetically altered plants could produce toxins and other substances that might harm birds and other animals .

Eventually , within the next few decades , agriculture will move off the soil and into biosynthetic industrial factories controlled by chemical and biotech companies . Never again will people know the joy of eating naturally produced , fresh foods . Hundreds of millions of farmers and other workers worldwide will lose their livelihoods . The hope of creating a human , sustainable agricultural system will be destroyed .

A .1 . B . Excerpt of an essay entitled “ Kill the Frankenstein Myth ” by Robert W . Tracinski , a senior writer for the Ayn Rand institute in Marina del Rey , California

In reality , anti biotech environmental activists claims against genetically modified foods are based not in science , but in a superstitious fear of science and technology . It is revealing that environmental activists have chosen to smear genetically modified foods with the term “ frankenfood ” , invoking Frankenstein , the classic horror story of a mad scientist who tampers with nature s secrets and unleashes a rampaging monster .

But this Frankenstein myth , and its theme of the dangers of science , has been thoroughly refuted in the nearly 200 years since it was first published . Science and technology have improved human life in countless ways , from the steam engine to the pasteurization of milk , from electrical power to antibiotics . And genetically modified foods are just the latest step in this march of progress .

Farmers have long modified the genetic makeup of their crops and livestock through selective breeding —choosing to breed the prize bull , for example , or planting seeds from the highest yielding stalks of wheat . But genetic engineering has made this process much easier and faster . For example , one popular variety of genetically engineered corn contains a gene taken from a bacteria , that gene produces a chemical toxic to caterpillars , giving the corn an inbuilt defense against harmful insects .

This new technology is already providing farmers with crops that bear higher yields , grow in drier climates , require fewer pesticides , and so on . The result has been bigger harvests and lower costs for American farmers . And scientists have also begun engineering plants that grow better under difficult conditions , such as drought —promising a new “ green revolution ” for the Third World .

Genetically modified foods are not merely safe —they are an enormous advance , and we should be applauding the heroes of science who invented them .

Part 2 : Writing an argumentative essay

Directions : Write a well organized academic essay (about 300 words) on the topic below . Your essay will be graded on content , organization , and appropriate use of English . You may refer to the reading passages while you are writing . Write your essay on the lined paper . You will have 45 minutes to write your essay .

A .2 . ESSAY TOPIC : Some people believe that genetically modified plants are dangerous to our health and to the environment . Others believe that genetic engineering is an important tool in feeding the world s population .

Which position do you support ?

DO NOT copy the original sentences or just write a summary of the reading passages .

Appendix B

Writing Test Paper (B)

Writing an argumentative essay

Directions : Write a well organized academic essay (about 300 words) on the topic below . Your essay will be graded on content , organization , and appropriate use of English . Write your essay on the lined paper . You

will have 45 minutes to write your essay .

B .1 . ESSAY TOPIC :Some people believe that genetically modified plants are dangerous to our health and to the environment . Others believe that genetic engineering is an important tool in feeding the world s population .

Which position do you support ?Give reasons for your answer .

You should use your own ideas ,knowledge and experience and support your arguments with examples and relevant evidence .

Appendix C

Rating Scale

Content	Organization	Accuracy	Vocabulary
<p>9 —10</p> <p>The treatment of the assignment completely fulfills the task expectations and the topic is addressed thoroughly .</p> <p>Uses a wide range of academic vocabulary .</p>	<p>9 —10</p> <p>Clear and appropriate organizational plan .</p> <p>Fully developed evidence for generalizations and supporting ideas / arguments is provided in a relevant and credible way .</p> <p>Connections between and within paragraphs are made through effective and varied use of transitions and other cohesive devices .</p>	<p>9 —10</p> <p>The essay is clearly written with few errors ;errors do not interfere with comprehension .</p> <p>Effective introduction and conclusion .</p> <p>Word choices are accurate and appropriate .</p>	<p>9 —10</p> <p>Uses a variety of sentence types accurately .</p> <p>Includes consistently accurate word forms and verb tenses .</p>
<p>7 —8</p> <p>The treatment of the assignment fulfills the task expectations competently and the topic is addressed clearly .</p> <p>Evidence for generalizations and supporting ideas / arguments is provided in a relevant and credible way .</p>	<p>7 —8</p> <p>Clear organizational plan .</p> <p>Satisfactory introduction and conclusion .</p>	<p>7 —8</p> <p>The essay is clearly written but contains some errors that do not interfere with comprehension .</p> <p>The essay may contain some errors in word choice , word form , verb tenses , and complementation .</p>	<p>7 —8</p> <p>The essay uses a variety of sentence types .</p> <p>Good range of vocabulary used with at most a few lapses in register .</p>

Content	Organization	Accuracy	Vocabulary
<p>5 —6</p> <p>The treatment of the assignment minimally fulfills the task expectations ; some aspects of the task may be slighted .</p> <p>Some relevant and credible evidence for generalizations and supporting ideas / arguments is provided .</p>	<p>Satisfactory connections between and within paragraphs using transitions and other cohesive devices .</p> <p>5 —6</p> <p>Adequate but simplistic organizational plan .</p> <p>Introduction and conclusion present but may be brief .</p>	<p>5 —6</p> <p>Is generally comprehensible but contains some errors that distract the reader ;at most a few errors interfere with comprehension .</p> <p>The essay may contain several errors in word choice , word form , verb tenses , and complementation .</p>	<p>5 —6</p> <p>Somewhat limited range of sentence types ; may avoid complex structures .</p> <p>Somewhat limited range of vocabulary .</p>
<p>3 —4</p> <p>The treatment of the assignment only partially fulfills the task expectations and the topic is not always addressed clearly .</p> <p>Evidence for generalizations and supporting ideas / arguments is insufficient and /or irrelevant .</p>	<p>Connections between and within paragraphs occasionally missing .</p> <p>3 —4</p> <p>Organizational plan hard to follow .</p> <p>Introduction and conclusion may be missing or inadequate .</p>	<p>3 —4</p> <p>Contains many errors ; some errors may interfere with comprehension .</p> <p>Includes many errors in word choices , forms , word forms , verb tenses and complementation .</p>	<p>3 —4</p> <p>Uses a limited number of sentence types .</p> <p>Vocabulary limited .</p>
<p>1 —2</p> <p>The treatment of the assignment fails to fulfill the task expectations and the paper lacks focus and development .</p>	<p>Connections between and within paragraphs frequently missing .</p> <p>1 —2</p> <p>No apparent organizational plan .</p>	<p>1 —2</p> <p>Contains numerous errors that interfere with comprehension .</p>	<p>1 —2</p> <p>Uses simple and repetitive vocabulary that may not be appropriate for academic writing .</p>

Content	Organization	Accuracy	Vocabulary
Evidence for generalizations and supporting ideas / arguments is insufficient and /or irrelevant .	Introduction and conclusion missing or clearly inappropriate . Few connections made between and within paragraphs .	Includes many errors in word choices , forms , word forms , verb tenses and complementation .	Does not vary sentence types sufficiently .

(...continued from p .109)

Mackey , A . & J . Philp . 1998 . Conversational interaction and second language development : Recast , response and red herrings . *The Modern Language Journal* 82 338-56 .

Nabei , T . & M . Swain . 2002 . Learner awareness of recasts in classroom interaction : A case study of an adult EFL student's second language learning . *Language Awareness* 11 44 .

Nicholas , H . , P . M . Lightbown & N . Spada . 2001 . Recasts as feedback to language learners . *Language Learning* 51 719-758 .

Norris , J . M . & L . Ortega . 2000 . Effectiveness of L2 instruction : A research synthesis and quantitative meta-analysis . *Language Learning* 50 417-528 .

Oliver , R . 1995 . Negative feedback in child NS -NNS conversation . *Studies in Second Language Acquisition* 17 : 459-481 .

Panova , I . & R . Lyster . 2002 . Patterns of corrective feedback and uptake in an adult ESL classroom . *TESOL Quarterly* 36 /4 573-592 .

Pica , T . 1994 . Research on negotiation : What does it real about second language learning conditions , processes and outcomes ? *Language Learning* 44 :493-527 .

葛现茹、高玉英 2005 关于师生对纠正性反馈的态度的调查 《渝西学院学报》第4 期。

何莲珍、王敏 2004 交际课堂中的形式教学——国外近期研究综述 《外语与外语教学》第1 期。

胡健、徐宏亮 2007 反馈语的特征与功能 《安徽大学学报》第2 期。

施光 2005 英语课堂中的教师纠错与学生接纳 《外国语言文学》第4 期。

宋金元、蔡小玲 教师反馈语对学生英语习得的影响 《嘉兴学院学报》第4 期。

孙燕青 2005a 第二语言学习中反馈 《心理科学进展》第2 期。

孙燕青 2005b 重述 :第二语言学习中的重要反馈方式 《心理发展与教育》第4 期。

韦静 2006 第二语言课堂教学中的教师纠正反馈语 《淄博师专学报》第2 期。

胥国红 2006 大学英语教师课堂反馈的功能研究 《西安外国语学院学报》第4 期。

赵晨 2005 不同水平英语教学中的教师纠正反馈语——一项基于语料库的研究 ,《解放军外国语学院学报》第3 期。