Advances in Language and Literary Studies
ISSN: 2203-4714
Vol. 5 No. 1; February 2014

The Impact of Time-Series Diagnostic Tests on the Writing Ability of Iranian EFL learners

Bahareh Molazem Atashgahi
E-mail: bmolazem@yahoo.com

Doi:10.7575/aiac.alls.v.5n.1p.146
URL: http://dx.doi.org/10.7575/aiac.alls.v.5n.1p.146

Received: 08/01/2014
Accepted: 26/02/2014

Abstract

This study aimed to show whether administering a battery of time-series diagnostic tests (screening) has any impact on Iranian EFL learners' writing ability. The study was conducted on the intermediate EFL learners at Islamic Azad University North Tehran branch. The researcher administered a homogenizing test in order to exclude the exceptional scores, among all the testers, only those whose scores were nearly within one standard deviation above or below the mean were selected as the participants of this study. After the assignment of the participants to the control and experimental groups- 30 students in each group- they were asked to write five-paragraph-essays on two topics. Such a pretest was given to both groups to test their initial writing ability. Once scoring of the students' writings (five-paragraph essay) was finished the two means of the groups were calculated and compared with each other through the t-test analysis. The result demonstrated that there was no statistically significant difference between those two groups regarding the variable under investigation. Four sets of diagnostic tests were given to the experimental group every two weeks and after each test both the result of the exam and suitable feedback regarding students' errors were given to them by the teacher, while the Current-Traditional Rhetoric method was administered in the control group.

In the posttest which was run after giving the treatment and placebo to experimental group and control group respectively, students took another writing test with the same characteristics in administration, topics and scoring as the one in pretest. Thereafter, the significance of the difference between the obtained means of experimental and control groups in the posttest was determined through the t-test. The result of the t-test analysis indicated a significant difference between the two groups which consequently rejected the null hypothesis of the study.

Therefore, any significant difference between the performance of experimental and control groups were attributed to the effectiveness of treatment which in this study was a set of parallel form diagnostic tests and the related feedback which was given by the teacher. Two matched t-test were also calculated to determine whether students in two groups had any improvements from the pretest to posttest or not.

Keywords: time-series, writing ability, feedback, diagnostic writing test

1. Introduction

Many scholars including (Harris 1969, Sako 1969, Wilkinson 1980, Madsen 1983, McDonough 1985) asserted that there are many elements to be considered in writing. These factors include: form, content, vocabulary, grammatical accuracy, penmanship, speed, mechanics, relevance, elaboration, originality, diction, layout, coherence, cohesion, unity, organization, and logic. Harris (1962) stated "writing as a complex skill involves the simultaneous practice of a number of very different abilities, some of which are never fully achieved by many students even in their native language" (p.68). He considered content, form, grammar, style and mechanics as the components of writing. In this regard, Madsen (1983) enumerated a number of different components and skills to be tested in writing. For Sako (1972) vocabulary, structure, accuracy and speed of script writing, spelling, punctuation, content and organization of material are the elements of writing. Meanwhile, MC- Donugh (1985) suggested grammar, coherence, relevance, and the structure of the argument as the attributes of a written task.

The attention to EFL writing has led to challenges among teachers for finding the most appropriate way of teaching writing in language classrooms. Writing as (Hilton and Hyder 1995) defined "is conveying our message in words through which we express our thoughts, ideas, questions, remarks, etc." (p.17).

From time to time, teachers may take an interest in assessing the strengths and weaknesses of each individual student in terms of the instructional objectives for the purpose of correcting an individual's deficiencies "before it is too late" (Brown 2005). To that end, diagnostic decisions are typically made at the beginning or middle of the term and are aimed at fostering achievement by promoting strengths and eliminating the weaknesses of individual students.

Naturally the primary concern of the teacher must be the entire group of students collectively, but some attention can also be given to each individual student (Brown, 2005). While diagnostic decisions are definitely related to

achievement, diagnostic testing often requires more detailed information about which specific objectives students can already do well and which they still need to work on.

Most teachers utilize direct assessment to measure the writing ability of their students as a holistic measure since it focuses on convention, linguistic and rhetorical knowledge of writing. However, teachers are never completely satisfied with the progression their students make although they put a lot of burden on their shoulders during teaching and scoring procedure.

According to (Alderson, Clapham et al. 1995) in a process of teaching a new language, especially in writing skill there always has been a lack of criteria to apply for screening the students during a course of study. The reason might be due to fact that normalization of the use of a series of diagnostic tests for the purpose of screening students has not yet occurred in most language classrooms. Teachers always limit themselves to provide students with feedbacks on their compositions that might be ignored by the majority of students.

There is no doubt that writing skill is the most difficult skill for L2 learners to master. The difficulty lies not only in generating and organizing ideas, but also in translating these ideas in to readable text. With so many conflicting theories around and so many implementation factors to consider, teaching a course in writing would be a daunting task. Therefore, it is important to know how to treat learner's errors (Richards and Renandya 2002).

One of the concerns of EFL teachers is to help students to develop the ability to produce correct and acceptable compositions. But there are very few studies that concern the manner in which teachers assess their students' foreign language skills whilst in the process of teaching and learning. General assessment studies on teacher behavior in language classrooms have shown that teachers spend a relatively small amount of time assessing individual student performance in order to diagnose their weaknesses (Edelenbos and Jong 2003).

According to (Truscott 1996) unfortunately, many teachers consider error correction in writing as just letting candidates receive a large amount of support in terms of feedback (mostly written corrective feedback) on their produced piece of writing. Truscott (1996) claimed that using this method for ESL (English as a second language) writers is ineffective.

On the other hand, providing students with an organized and relevant feedback through the use of diagnostic tests whose purpose is to find and then focus on learners' real deficiencies and needs would be beneficial.

In general, the primary goal of traditional educational tests is to make inferences about an individual test taker's general ability with reference to other test takers in the normative group (Brown and Hudson 2002). Such traditional testing has been criticized for not providing diagnostic information to inform students of their strengths and weaknesses in a specific academic domain (Snow and Lohman 1988). As standardized tests are thus being increasingly recognized as unsatisfactory (Mislevy, Almond et al. 2004), testing communities have called for more diagnostic information for guiding learning, improving instruction, and evaluating students' progress.

Teachers are not only interested in taking program-level decisions but are most interested in classroom tests. A test is considered influential when it can help teachers to find a good direction which shows them what to teach. In this respect, Kinsena (1985) states that "without a fundamental awareness of our performances, it is easy to believe that the way we study and learn is the most efficient way and consequently help teachers to diagnose some of their students' problem" (p.32).

2. Method

2.1 Participants

The participants of the study were 60 English translation students who were studying at Islamic Azad University. All of them were English translation students at B.A. level. Participants who were all junior students had already enrolled in essay writing class. The participants, both female and male students, aged from 19 to 26. Gender and age of the participants were not taken into consideration in this study. They were all Persian native speakers who were learning English as a foreign language. It is worth mentioning that the instructor was the same in both classes. Each group consisted of 30 participants.

2.2 Instrumentation

The instruments which were employed in this study included a homogenizing test to assure homogeneity of the participants of the study, tests of writing which were served as the pretest and post test and a rating scale used for giving scores by the raters.

2.2.1 Homogenizing test

To minimize the individual differences among the participants and to ensure homogeneity, a paper-based TOEFL test, which was a standard test, was used as a reliable and standard criterion. The test was taken from *Longman Preparation Course for the TOEFL by (Phillips 2005).* The first part consisted of 40 items, including 30 instruction and written expression items along with 10 reading comprehension questions, all in multiple choice formats. The second part, however, was TWE (Test of Written English) in which students were supposed to write a five-paragraph essay on a given topic.

2.2.2Pretest, posttest

The pretest and posttest were consisted of two writing topics for each test. Participants were supposed to write two five-paragraph essays with the length of not less than 150 words and within a time limit of 60 minutes. The topics were adopted from www.ets.org which is the official website of TOEFL organization.

2.2.3 Rating scale

After the administration of the test, the essays were scored by two raters. Both raters used Jacobs et al. s' writing template (1981) called "ESL Composition Scoring Profile".(Jacobs 1981).

2.3 Material

Four sets of diagnostic tests, each consisted of 30 questions were used as the treatment in experimental group. To name few, in each set of these tests, some constructs as comma spliced, dangling structure, conciseness, content as well as coherence along with 25 other constructs were tested. A complete list of these constructs is presented in a table of specification. Four sets of diagnostic tests were adopted from Prenhall website, which is the website of Pearson Higher Education.

2.4 Procedure

As it was mentioned before the aim of this study was to show whether screening learners trough a battery of time-series diagnostic test had any impact on the Iranian EFL learner's writing ability.

The study was conducted on 60 B.A. English translation students of Islamic Azad University North-Tehran Branch. First of all, a homogenizing test was conducted both in experimental group and control group in the first session. After analyzing the result of this test, the number of participants (N=80) turned out to be 60 since nearly those whose scores were within one standard deviation above or below the mean were selected as the participants of this study. Each group consisted of 30 students who were mostly female.
After that, in the second session, the participants in the control and experimental groups were asked to write on certain topics, which served to be the pretest to test their initial writing ability. After the administration of the test, the essays were scored by two raters. Both raters used Jacobs et al. s' writing template (1981) called "ESL Composition Scoring Profile".

Jacobs' et al. (1981) criteria have been researched and found to have construct validity. Construct validity, is the degree to which the scoring instrument is able to distinguish among abilities in what it sets out to measure, and is usually referred to in theoretical terms; in this case, the theoretical construct is ''essay writing ability'' which the instrument aims to measure.

In the following sessions both experimental and control groups were taught essay writing through current-traditional rhetoric approach at an advanced level. They both worked on different types of essay as well as the mechanics of writing.

The experimental group, however, was exposed to the treatment, which was four sets of time series (screening) diagnostic tests along with their related feedbacks, from the third session.

The first diagnostic test was administered in the third session. The papers were scored by the teacher and weaknesses of each individual were diagnosed. Within the next two sessions the teacher gave appropriate and related feedback to the class in accordance with the weaknesses of the majority. Finding out which points students were weakest at was the responsibility of the teacher himself. The teacher provided the students with related feedback by either teaching those points in the class or through an oral discussion. For example for the following question, according to teachers' report, 18 students chose letter c and other 12 students selected other letters except letter A which was the correct answer.

Q7: It was her who won the election _ for a new union representative.
A B C D

As the table of specification shows the construct under investigation in this item was pronoun case. Accordingly, this point became one of the major concerns of the teacher in the following two sessions to be taught.

The second diagnostic test was held in sixth session, and again after two sessions of instruction the third and fourth tests were held in tenth and thirteenth sessions respectively. No need to mention that the teacher went through the same process of diagnosing and teaching in the following sessions of the treatment.

It is good to mention that the students were aware that every two sessions they were having a test.

Providing any types of feedbacks was completely dependent on the result of each test. In this way not only the result of the test was given to students, but also the teacher made them aware of the problematic areas and enabled them to see how much of the progression they made.

On the other hand, the control group spent 13 sessions just practicing the Current-Traditional Rhetoric approach with no focus on the problematic areas through giving relevant feedback. The main focus of the class went to getting students to know the types of essays and the mechanics of writing while no extra attention was given to the errors they had in their essays and no attempt to solve them.

At the end, in fifteenth session students in both experimental group and control group were asked to write two other essays, a posttest, with the same characteristics in administrating, scoring, but different topics. It is necessary to point out that the posttest was the same as their final exam. So it can be claimed that they did their best since to them it was a high stake exam.

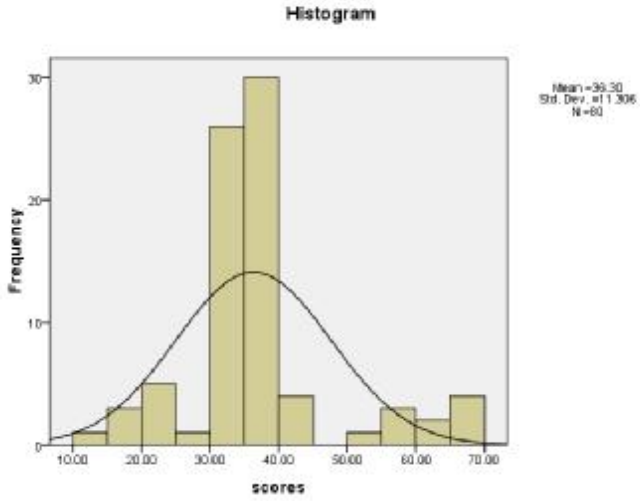
3. Results

Table 1 shows the descriptive statistic results of the homogenizing test. As the table shows the skewness value turned out to be .181 and the standard error of skewness was .296. The subdivision of skewness by standard error of skewness turned out to be .611 which is between -1.96 and +1.96. Therefore, it can be concluded that the distribution is normal.

Table 1. Descriptive statistics of general proficiency test

N	Minimum	Maximum	Mean	Std. Error of Mean	Std. Deviation	Variance	Skewness	Std. Error of Skewness
80	12.00	69.00	36.30	1.26	1.13	127.833	.181	.269

The following figure shows the normal distribution after homogenizing test



An independent t-test was run to compare the mean scores of the experimental and control groups on the pretest. As displayed in Table 2, two groups were not significantly different since at .05 level of significance for 58 degrees of freedom, the significance 2-tailed was .917 which is higher than .005.

Table 2. Independent Samples Test

		Levene's Test for Equality of Variance		t-test for equality of means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% interval Difference Lower	Confidence of the Upper
PRETEST	Equal variances assumed	.370	.545	.104	58	.917	.07842	.75229	-1.42745	1.58429
	Equal variances not assumed			.105	57.947	.917	.07842	.74980	-1.42249	1.57933

Thus it can be claimed that the two groups were homogenous in terms of their writing ability prior to the administration of time series (screening) diagnostic tests. The descriptive statistics for the two groups are displayed in table 3.

Table 3. Descriptive Statistics Pretest

	GROUP	N	Mean	Std. Deviation	Std. Error Mean
PRETEST	Experimental	30	12.9655	2.76112	.51273
	Control	30	12.8871	3.04606	.54709

In order to test the null hypothesis, an independent t-test was run to compare the mean scores of the experimental and control groups on the posttest.

Table 4. Independent t-test Posttest

		Levene's Test for Equality of Variance		t-test for equality of means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence interval of the Difference Lower	Upper
POSTTEST	Equal variances assumed	.637	.428	3.605	58	.001	2.38543	.66169	1.06092	3.70994
	Equal variances not assumed			3.616	57.966	.001	2.38543	.65963	1.06502	3.70584

As displayed in table 4, two groups were significantly different since at .05 level of significance for 58 degrees of freedom the significance 2-tailed was .001, which is lower than .005.Thus it can be claimed that there was a significant difference between the two groups' mean scores on the posttest.

As shown in table 5, the experimental group with a mean of 14.7241 outperformed the control group whose mean was 12.3387. Based on these results it can be concluded that the null-hypothesis that administering time series (screening) diagnostic tests does not optimize the writing ability of the Iranian EFL learners is rejected. The administration of time series (screening) diagnostic tests had improved the writing ability of the experimental group. The descriptive statistics for the two groups are displayed in table 5.

Table 5. Descriptive Statistics Posttest

	GROUP	N	Mean	Std. Deviation	Std. Error Mean
POSTTEST	Experimental	30	14.7241	2.43701	.45254
	Control	30	12.3387	2.67204	.47991

To prove that the experimental group outperformed the control group on the posttest does not guarantee that the control group has not improved on the posttest compared with its mean score on the pretest.

In order to make sure that there is no significant difference between the pretest and posttest mean scores of the control group, a paired t-test was run. The statistical results are shown in table6.

Table 6. Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	precontrol	12.8333	30	3.08314	.56290
	postcontrol	12.2167	30	2.62837	.47987

As displayed in table 4.7, there was no significant change regarding writing ability of the control group, since the post-independent test at significance level of .05 at 29 degrees of freedom the significance 2-tailed was .335 which was higher than .005. Therefore the existing difference is not statistically significant.

Table 7. Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	precontrol - postcontrol	.61667	1.04232	.20856	.19012	1.04321	.93	29	.355

A paired sample statistic was also calculated in order to make a comparison within the experimental group. The means of this group in both pretest and posttest were compared with each other.

As displayed in Table 4.8, the mean score of experimental group in pretest was compared with its mean score in posttest. The result was that the experimental group with the mean of 14.7667 in the posttest outperformed the pretest with the mean of 13.0167.

Table 8. Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	preex	13.0167	30	2.72752	.49797
	postex	14.7667	30	2.40593	.43926

As displayed in table 9 there was a significant difference between the performance of the experimental group on the pretest and posttest. The post independent t-test at the significant level of .05 for 29 degrees of freedom, the significance 2-tailed was .000 which was lower than .005 and therefore the existing difference is statistically significant. Thus, it can be claimed that there was a significant difference between the pretest and posttest mean scores of the experimental group. As a result the improvement in the experimental group can be attributed to the treatment, which in this study was a set of time series diagnostic tests.

Table 9. Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	preex - postex	-1.75000	1.32450	.24182	-2.24458	-1.25542	-7.237	29	.000

3.1 Inter-rater reliability

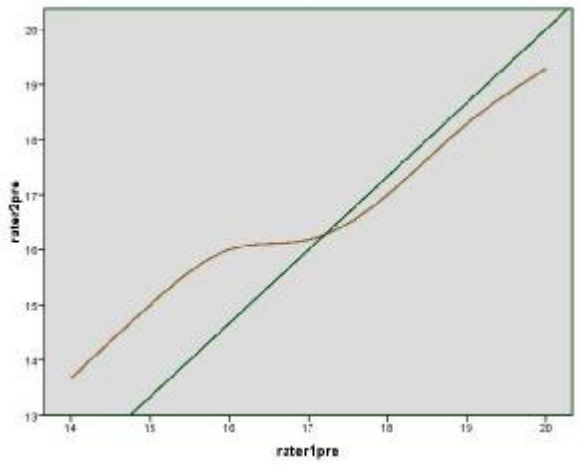
As displayed in Table 10, there is a high degree of consistency in the judgment of the raters in pretest.

r = .80

Table 10. Inter-rater Reliability

		RAT2
RAT1	Pearson Correlation	**.804 (*)**
	Sig. (2-tailed)	**.000**
	N	**60**

*. Correlation is significant at the 0.05 level (2-tailed).

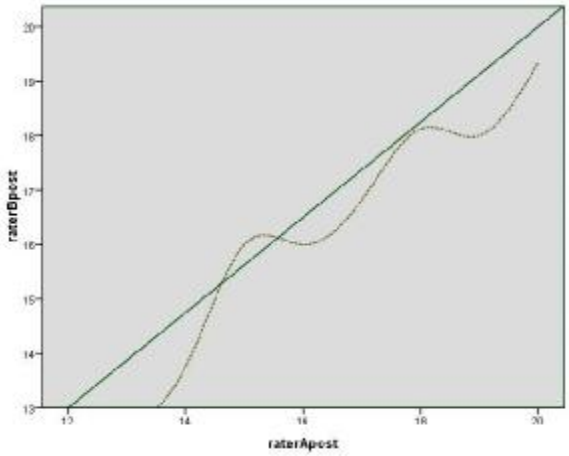


As displayed in Table 11 there is a high degree of consistency in the judgment of the raters.

r = .84

		RAT2
RAT1	Pearson Correlation	**.842(*)**
	Sig. (2-tailed)	**.000**
	N	**60**

*. Correlation is significant at the 0.05 level (2-tailed).



4. Discussion

In this study, the researcher concluded that administering time series diagnostic (screening) tests had statistically significant impact on the improvement of the writing ability of Iranian intermediate level EFL learners. The experience of using time-series diagnostic tests for university-level students was helpful in the way that both students and teachers were exposed to the real writing ability which takes both strong and weak points of students into consideration. Screening, as shown in this research, requires that teachers have a constant eye on the progress of students to find where they have problems and to resolve them since they could not successfully commit individually. The teacher's initial role always is to diagnose which points to emphasize more and also provide feedback for.

The result of the current study supported the previous ones'(Alderson and Banerjee 2001, Ishii and Schmitt 2009) in terms of using diagnostic tests for making both teaching and learning more effective. For example, (Ishii and Schmitt 2009) investigated the effect of an integrated diagnostic test of vocabulary on the vocabulary learning of students and concluded that in this way, it is possible to diagnose any weak areas of the learners' vocabulary knowledge in advance and as a result a principled way of treating these deficiencies can be used.

Due to practical limitations in randomization the number of students participating in this study was relatively small (30 in experimental group and 30 in control group). Additionally, treatment had focused only on one skill, i.e. writing, while the participants my benefit from this method in other three language skills like speaking. Moreover, the study was not a longitudinal one since the whole took only one semester with 15 sessions.

If this method included in educational schedules, might empower the subjects with an insight into the language they are learning. In addition, it provides a means of self-monitoring different from what is common in writing classes. Unlike the other types of tests, the present diagnostic test has the potential to be far superior, because the weaknesses and inadequacies of individuals are caught and dealt with by the teacher. Perhaps the most effective use of this method is to report the performance level on each writing element to the teacher and each student so that they can decide how and where to most profitably invest their time and energy.

It would also be useful to report the average performance level for each class on each objective to the teacher along with indicative of which students have particular strengths or weaknesses of each objective (Brown 2005). For reaching to effective results, the instructor must view teaching as a process of developing and enhancing students' ability to learn. The instructor's role is not just to teach some preplanned materials, but to serve as a facilitator for learning by providing relevant feedback regarding students' weaknesses. This may also result in increasing confidence of students.

Another study may be done to explore the impact of the method on the improvement of the other language skills, listening, reading and speaking as well as language components with beginners and advanced learners. Promoting the motivation and attitudes of EFL learners toward the writing skill, the effect of the same method in a longitudinal research, and the role of gender on the performance of both groups may lead to different results.

Olshtain as cited by(Celce-Murcia 1991) believed that writing, as a communicative activity requires to be encouraged and nurtured during the language learners' course of study. While the most important thing during a course is to identify weak points of learners and to remove them, Alderson (2005) suggests that diagnostic tests should be used to identify strengths and weaknesses in learners' use of language and focus on specific elements rather than global abilities. In writing area, few diagnostic assessments are specifically designed for providing diagnostic feedback (Alderson, 2005; Gorin, 2007). Therefore, there is a great need for a diagnostic test that helps both teachers and learners to find out their source of errors then eliminate those problems.

References

Alderson, J. C. and J. Banerjee (2001). "Language testing and assessment (Part I)." *Language Teaching, 34*(04): 213-236.

Alderson, J. C., et al. (1995). Language Test Construction and Evaluation, Cambridge University Press.

Brown, J. D. (2005). Testing In Language Programs: A Comprehensive Guide To English Language Assessment, Prentice Hall Regents.

Brown, J. D. and T. Hudson (2002). Criterion-Referenced Language Testing, Cambridge University Press.

Celce-Murcia, M. (1991). "Grammar Pedagogy in Second and Foreign Language Teaching." *TESOL Quarterly, 25*(3): 459-480.

Edelenbos, p. and J. D. Jong (2003). "Vreemdetalenonderwijs in Nederland:Een situatieschets [Foreign language teaching in the Netherlands: a situational sketch]." Enschede: NaB-MVT.

Harris, D. P. (1969). Testing English as a second language, McGraw-Hill.

Hilton, C. and M. Hyder (1995). Getting to Grips with Writing, Golden Books Centre.

Ishii, T. and N. Schmitt (2009). "Developing an integrated diagnostic test of vocabulary size and depth." *RELC Journal, 40*(1): 5-22.

Jacobs, H. L. (1981). Testing ESL Composition: A Practical Approach, Newbury House.

Kinsena,L., (1985*).* The effect of mode-discourse on student writing performance: implications for policy. *Educational Evaluation and Policy Analysis, 8(2),*147–154.

Madsen, D. (1983). Successful dissertations and theses, Jossey-Bass.

McDonough, S. (1985). "Academic writing practice." *ELT Journal, 39*(4): 244-247.

Mislevy, R. J., et al. (2004). A Brief Introduction to Evidence-centered Design, National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation, Graduate School of Education & Information Studies, University of California, Los Angeles.

Phillips, D. (2005). Longman Preparation Course for the TOEFL Test: Next Generation IBT ; [with Answer Key], Pearson Education.

Richards, J. C. and W. A. Renandya (2002). Methodology in Language Teaching: An Anthology of Current Practice, Cambridge University Press.

Sako, S. (1969). "Writing Proficiency and Achievement Tests." *TESOL Quarterly, 3*(3): 237-249.

Snow, R. E. and D. E. Lohman (1988). Implications of Cognitive Psychology for Educational Measurement, Stanford University, CERAS.

Truscott, J. (1996). "The Case Against Grammar Correction in L2 Writing Classes." *Language Learning, 46*(2): 327-369.

Wilkinson, A. M. (1980). Assessing language development, Oxford University Press.