THE TOEIC TEST AND COMMUNICATIVE COMPETENCE:

Do Test Score Gains Correlate With Increased Competence?

a preliminary study

by

CYNTHIA R. CUNNINGHAM

A dissertation submitted to the
School of Humanities
of the University of Birmingham
in part fulfillment of the requirements
for the degree of

Master of Arts
in
Teaching English as a Foreign or Second Language (TEFL/TESL)

This dissertation consists of approximately 12,000 words.
Supervisor: Clare Hindley

Centre for English Language
Studies
Department of English
University of Birmingham
Edgebaston, Birmingham B15 2TT

September 30, 2002

Abstract

The TOEIC test is an internationally accepted, multiple-choice test of general English proficiency. It is marketed as an indirect yet highly reliable measure of non-native speakers' abilities to communicate in an English-speaking work environment. However, there is a lack of independent research into whether or not the TOEIC test does indeed measure communicative competence.

To explore this question, a direct assessment of listening, reading and writing abilities (TIC) was created and administered to a homogenous group of first year university students, paralleling TOEIC test dates. Results from both the entry and exit tests were analysed to determine if gains on the TOEIC test correlated with improved communicative competence, as measured by TIC. The findings were discussed in relation to the literature on the TOEIC test, testing and communicative competence.

Initial results suggest that a correlation neither exists between TOEIC test scores and communicative abilities, nor between TOEIC test score gains and improved communicative competence. Additional findings suggest that TOEIC test-preparation does not result in more accurate use of structure. Furthermore, it appears that the test is not an ideal discriminator of language abilities. Thus, its role as a placement test and as a measure of non-native speakers' English language abilities needs reappraisal.

ACKNOWLEDGMENTS

I am most grateful to my supervisor, Clare Hindley, for her insight and for the positive and constructive criticism she provided. As well, I truly appreciate the time she spent reading the various stages of this dissertation and with the speed in which she provided comments.

I am particularly indebted to Paul Moritoshi for agreeing to be the second-marker; for sharing resources, and for the emotional support and thoughtfulness he extended during this project.

I am most thankful to Kazuma Kawamura for his help in translating both the Japanese newspaper articles and the TIC test tasks completed in Japanese.

I would like to express my sincere thanks to Takashi Miura for his guidance on the earlier modules of the ODL program.

Finally, I would like to thank the students for completing the tests to the best of their abilities.

CONTENTS

LIST OF TABLES

CHAPTER 3

CHAPTER 4

CHAPTER 5

Chapter 1 INTRODUCTION

The Test of English for International Communication (the TOEIC) is a multiple-choice (M/C) proficiency test of listening and reading skills. Educational Testing Services (ETS) and the Chauncey Group International Ltd., respectively the developers and administrators of the TOEIC program, describe the test as a direct measure of those abilities and as an indirect measure of speaking and writing. They also claim a strong correlation between TOEIC scores and English communicative abilities (Woodford 1978; ETS' *The Reporter* #4; TOEIC Examinee Handbook 1996). Yet as the format includes neither a spoken nor a written discourse component, some feel that the claim is misleading for and/or misinterpreted by test users (Childs 1995; Gilfert 1995; Gilfert 1996; Hamp-Lyons 1998; Hamp-Lyons 1999; Hilke and Wadden 1999; Smith 2000).

Most available research is ETS published, leaving the results and interpretations open to criticisms of bias. Considering the influence TOEIC test results exert on the future of test-takers (for example see the TOEIC Newsletter *The Reporter* #5; #6; #11), the lack of independent research is disturbing.

This paper investigates the relationship between TOEIC test score gains and increased communicative competence through consideration of the following questions:

1. Do TOEIC score gains correlate with improved English communicative abilities, in both comprehension and production, as measured by a direct test of listening, reading and writing abilities?

2 Does TOEIC test preparation see results in both fluency and accuracy with regards to the use of English grammar and vocabulary?

To address the above questions, I designed a listening and reading/writing test (Test of Interactive Communication - TIC) to measure how much information an EFL (English as a Foreign Language) student could gather from both spoken and written texts, and how well they could convey this information in written form. TIC consists of five listening tasks and five reading tasks, which examinees complete in either short-form or extended-response. Four groups of first-year university students took TIC during the first week of the 2002 spring-term and again towards the end of the term, paralleling the TOEIC entry and exit test dates held in one university's English language program.

Score gains and losses on the TOEIC are correlated with the same on TIC to discern any significant relations. I believe the results will re-enforce the skepticism of a direct link between TOEIC test score gains and increases in communicative competence. Gains may be made on the TOEIC but similar gains in the students' use of grammar and vocabulary will not be realized on TIC. However, I believe the structure of the college's first-year English program will provide for increases in the students' listening abilities and to a lesser extent, fluency.

This paper proceeds by first describing the situation of EFL education in Japan and in one university in particular (Chapter 2). Chapter 3 will feature a review of literature pertaining to testing and the TOEIC test within a global context, communicative competence, second language acquisition (SLA) and the phenomenon of washback. Chapter 4 will provide an in-depth description of TIC and its administration, followed by a discussion in Chapter 5 of the results as they relate to the research questions, the implications of this research for the Japanese EFL classroom and to the extent possible, other non-native speaking English countries.

Chapter 2 ENGLISH AS A FOREIGN LANGUAGE IN JAPAN

2.1 The Tradition of Foreign Language Education

Japanese students receive six years of formal EFL instruction commencing their first year of junior-high school. The Ministry of Education's (*Mombusho)* aims are an appreciation of the English language, the development of positive attitudes towards and basic abilities in active and practical communication (*Mombusho* 1999, cited in Moritoshi *nd* :3). However, students generally spend these years studying vocabulary and learning English through *YAKUDOKU*, a form of grammar translation (Gorsuch 1998; Hino 1988; Leonard 1998; Moritoshi *op cit*), as part of their university entrance-exam preparation.

YAKUDOKU is a three-step process where students translate a text into Japanese word-for-word, reorder it to comply with Japanese syntax and then recode it using Japanese particles (Gorsuch *op cit*; Hino *op cit*). The translated text is used as a basis for content discussion or language analysis conducted in Japanese. As a result, students tend to have a high knowledge about English but low competency in using it (Brown and Yamashita 1995; Gorsuch 1998; Hino 1988).

Most universities require students to complete a first-year general-English language course to develop their English language communication skills. Class placement is based on a pre-term M/C proficiency test score. An exit test is also administered, with score gains or losses calculated into the students' overall grade. The tests of choice are the TOEIC and the TOEFL (Test of English as a Foreign Language); both are products of ETS.

2.2 The Role of Proficiency Tests

One component of these first-year English courses is TOEIC/TOEFL test-preparation, using texts that develop test-taking skills and focus on discrete points of language (Hamp-Lyons 1998; Hilke and Wadden 1997). The reason for this emphasis with the TOEIC would appear to be two-fold. At the administrative and departmental levels, TOEIC score gains are deemed indicative of students' English communicative abilities which may be needed for future employment, while inclusion of the test in the curriculum is considered to motivate students in their language studies. For students, high scores are intrinsic to finding employment after graduation (Asahi Shimbun 2000; Sankei Shimbun 1999).

Many companies consider TOEIC test benchmark scores during the application process. ETS has published functional descriptions for specific TOEIC score-bands (TOEIC Examinee Handbook 1996; TOEIC Can-Do Guide 1998; TOEIC UserGuide 1999), which are accepted by various companies (Tenth TOEIC Client Survey 1999). However, ETS does recommend establishing benchmarks according to individual company needs (TOEIC UserGuide 1999). Universities also adopt these descriptions to establish students' goals, creating a high-stakes situation for the learners and arguably inappropriately using the test.

2.3 Proficiency Testing at One Japanese University

This section will describe how one Japanese university has incorporated the TOEIC test into its general-English language program. An in-depth description of the test can be found in Section 3.1.1.

The college offers four-year programs in Business Administration, Economics and

Accounting, as well as post-graduate studies in Economics. Its freshman-English course has four 1-hour classes a week, approximating 48 classes per term. Innovations to the 2002-2003 curriculum highlight an increased emphasis on the TOEIC test: graduation requirements include a TOEIC score of 425 points, regardless of how students fare in the course itself. Students who do not obtain a score of 450 during their freshman year must take additional English courses. While details have not yet been announced for these supplementary classes, the first year is comprised of two TOEIC test-preparation/business-English classes and two oral English classes.

The college uses the TOEIC as both an entry and exit test, with score gains and losses worth 30% of the students' English grade. The first term's exit test score becomes the students' base score for the second term. In many cases, expected gains approximate 80 points per term.

ETS recommends a minimum of 100 hours of instruction before significant improvement in either language ability or TOEIC scores can be expected. The university's language program obviously falls short of this, yet demands substantial score gains of the students. It is an issue that the author has previously discussed (Cunningham 2003).

The TOEIC (see Section 3.1.1 for a full description) is a norm-referenced test (NRT), meaning students' scores are compared with other students' rather than with previous individual efforts. As NRTs have no relation whatever to the progress made by the student within the goals of the language course, Gorsuch (1997) and Brown (1995) argue they should not be used as achievement tests. Unfortunately, this is one of the applications of the TOEIC within the college (Cunningham, *op cit*).

ETS states that as a proficiency test, the TOEIC cannot be studied for; instead,

TOEIC scores increase as the English level of the examinee increases (TOEIC UserGuide 1999). At TOEIC seminars (The TOEIC Steering Company (a); *ibid* (b)), key ETS members clearly state that the TOEIC tests neither writing nor speaking, and the key is to improve one's English ability. Conversely, TOEIC publications make reference to test-preparation books and CD-ROMS as a source of study. Since 2000, ETS has published its own preparation textbook. Such conflicting information from source-published material has the potential to cause confusion or lead to misunderstandings regarding what the test can or cannot do and for what it was designed to be used. This conflict has been picked-up even by supporters of the test (Hilke and Wadden 1997; *ibid* 1999).

While this illustrates how one university may be inappropriately using the TOEIC, similar applications may be found in other organizations world wide, as is demonstrated in the following chapter.

Chapter 3 LITERATURE REVIEW

In this chapter, I conduct a literature review as it pertains to the TOEIC test and testing, communicative competence, theories of SLA and washback, and the relation between them.

3.1 The Test of English for International Communication

3.1.1 Description of the TOEIC Test

The TOEIC test consists of two parts, each with 100 questions. The listening comprehension section has four sub-tests, and lasts 45 minutes; the structure/reading comprehension section has three sub-tests, and lasts 75 minutes. A description of the test-items may be found in Table 3.1; example test-items from the ETS published preparation-text (Arbogast *et al* 2000) are provided in Appendix 1.

An examinee's score is based on the number of questions correctly answered, in increments of five points; both the listening and reading sections of the test are thus graded on a scale of 5 – 495 for a combined maximum score of 990 points. There is no penalty for wrong answers and examinees are encouraged to guess as this increases their potential for a higher score (*The Reporter* #4 1990), a peculiar recommendation if the test is to be an accurate, reliable and valid measure of English language abilities.

The test was created for NNSs of English who use or expect to use English for communication in their work. As such, the test-items' context is business-oriented, including settings and situations such as general business, manufacturing, finance, corporate development, travel, entertainment and health.

Table 3.1 Description of the TOEIC test items (adapted from the TOEIC Examinee Handbook, The Chauncey Group Ltd. 1996)

PART 1 – Listening Comprehension, pictures	For each question, there is a photograph in the test booklet; the examinee hears 4 short statements, and must choose the statement which best describes the photograph. The statements are not printed in the test booklet. Settings include offices, street scenes, restaurants, laboratories, *etc.*
PART 2 – Listening comprehension, question - response.	For each item, the examinee hears a question, followed by 3 responses. Neither the question nor the answer is printed in the test booklet. The examinee must choose the correct answer to the question.
PART 3 – Listening comprehension, short conversation	For each question, the examinee hears a 3-part exchange; they then read a short question and 4 possible answers to the question. Situations encompass work-related discussions, meeting and business trip plans or schedules, requests for information at airports or train stations, *etc*.
PART 4 – Listening comprehension, short talks	For each item, the examinee listens to a short talk; printed in the exam booklet are 2 or more questions related to the talk. Each question has 4 possible answers. Talks include public announcements, news reports, meeting discussions, public service bulletins, and commercials.
PART 5 – Reading comprehension, incomplete sentences	Each question is an incomplete sentence. Four options to complete the sentence are listed beneath it. The examinee must choose the correct word or phrase that completes the sentence. Missing items are either based in word meaning or form, collocations or grammatical structure.
PART 6 – Reading comprehension, error recognition	Each item has four words or phrases underlined. The examinee must identify which of the four is incorrect. The examinee need not correct the sentence, only identify which item needs to be corrected. Types of errors are similar to the missing items in Part 5.
PART 7 – Reading comprehension	This part of the test is comprised of a variety of reading material taken from a business context or everyday affairs, such as notices, letters, forms, advertisements, newspapers, schedules, forms and applications. For the questions related to each text, the examinee must choose the correct responses from a choice of 4. The correct answer is based on what is stated or implied in the text.

As has been pointed out, the test-items themselves are not in context and thus the test is operating within an "artificial reality" (Gilfert 1995:84; Gilfert 1996). While examinees may listen to a telephone message, they need not take a message nor leave one. However as the next section indicates, this artificiality is interpreted as virtual reality.

3.1.2 Applications of the TOEIC Test

Since its inception in 1979, the TOEIC test quickly gained prominence within the field of EFL/ESL (English as a Second Language) education. It has been lauded as an efficient and reliable test of EFL/ESL communicative abilities, as the following statements corroborate:

1. "Assuring that SLOOC employees had good communication ability in English would not be easy without a reliable measure. Early in 1984, SLOOC began using TOEIC." (*The Reporter* #1:1).
2. "So many Thais study English but can't communicate in it. TOEIC helps us find those who can…." (*The Reporter* #2:1).
3. "Why do you use TOEIC? …it is useful to know if staff can speak as well as is needed to fulfill their responsibilities in a particular position." (*ibid*:3).
4. "To measure the effectiveness of the PG&G program and to report quantifiable results to its client, The Experiment uses the TOEIC test. Accountability is built into the program." (*The Reporter* #4:1).
5. "Academic institutes should be encouraged to use the TOEIC to measure students' general English ability." (Eotvos University in *The Reporter* #4:4).
6. "TOEIC is testing exactly the language that companies want their employees to be able to control." (*The Reporter* #8:3).

Results from the test are used by organizations "to make significant personnel decisions" (The TOEIC UserGuide 1999:6), for example evaluating personnel, selecting candidates for training conducted in English, for recruiting, promoting and sending employees overseas, and for identifying "employees who required further English language training, to set learning goals, and to monitor their progress" (*ibid* 1999:7; Asahi Shimbun 2000; Nikkei Shimbun 2000; Mainichi Shimbun 2000). Scores are also used by language schools for placement and evaluation and by universities as a graduation requirement (The TOEIC UserGuide 1999). It is very much a high-stakes test,

but as Childs (1995) and Schneider (2001) argue, incorrectly so due to the low reliability of individual test scores.

As was stated in the Introduction, little independent research has been conducted to verify the test's claim of being a reliable and valid indicator of communicative abilities, yet the above quotes indicate many organisations use the test in this capacity. This misperception of the test's ability is understandable when one encounters powerful statements in ETS published material. For example:

> "Research…has established that scores on the TOEIC correlate highly with direct measures of speaking proficiency. This means that …(you) can interpret TOEIC scores as indirect measures of active skills with a relatively high degree of reliability…Informal impressions of oral production, however, are highly subjective and unreliable…."
>
> (*The Reporter* 1990:3)

> "The TOEIC Test measures the everyday English skills of people working in an international environment. Test scores indicate how well people can communicate in English with others in the global workplace."
>
> (The TOEIC Technical Manual, electronic edition)

While this may indicate the level of acceptance garnered by the TOEIC, it does not explain the demand for the test.

3.1.3 The Need for the TOEIC

Since 1963, *Mombusho* has encouraged the use of the *Eiken* Test (The Society for Testing English Proficiency – STEP), a five-level test that measures candidates' aural abilities and structural knowledge via an M/C test, their written abilities via a written composition and their speaking abilities via an oral interview (the *Eiken* Website).

However, due to the problems employees posted overseas were having in communicating in English and with the perceived unsuitability of the TOEFL' academic nature for business people, Japan's Ministry of International Trade and Industry (MITI) requested ETS to develop a test specifically for business people (for example, TOEIC History and Status *nd*; TOEIC UserGuide 1999).

ETS was to

> "…develop highly valid and reliable measures of real-life reading and listening skills and to the extent possible, indirect measures of speaking and writing" (Woodford 1978:2)

as a departure from the traditional yet typical foreign language test of reading and grammar. MITI also wanted a test which would positively influence English language instruction (washback) in Japan and worldwide (*ibid).*

This request is interesting for three reasons. First of all, a direct and reliable measure already existed in the STEP test, raising questions as to why an indirect, M/C test was perceived as being a more accurate assessment of communicative abilities than a direct test. Perhaps the difficulties experienced by the overseas employees reflected their high structural competence but weak pragmatic abilities in English, a problem not necessarily solved through a different test-format.

Furthermore, real-life interaction does not consist of multiple-choice options, thus negating claims of being a 'measure of real-life reading and listening skills' (Woodford, *op cit*). Arguably, more items encompassing a wider domain of language use may be covered in such a format but it is still only a predictor of abilities, at best.

This leads into the second point. Woodford states

> "It is often the case…that the ability to understand spoken language together with an ability to manipulate grammatical structures and vocabulary, even on paper, can give an indication of an examinee's ability to speak (*ibid*:4).

However, researchers doubt the validity of such claims (Rutherford 1987; Weir 1990; Larsen-Freeman and Long 1991), even questioning the validity of using M/C formats for elicitation tests (Lewkowicz 2000; Morrow 1979; Richards 1980, cited in Larsen-Freeman and Long 1991:42; Weir 1990; Wu 1998).

Thirdly, ETS states that the advantage of the TOEIC over the TOEFL is that the TOEFL is academic in nature, while the TOEIC allows candidates to 'demonstrate their ability to use English in the workplace…' (TOEIC UserGuide 1999:8). Yet referring back to Table 3.1, the TOEIC' format does not require examinees to demonstrate an ability to *use* the language; neither are they required to *manipulate* it. Indeed, a comparison of the two tests (Table 3.2) indicates more similarities than differences. The main difference would be the context (Gilfert 1995; Hemingway 1999).

Table 3.2 A Comparison of the TOEIC test and the TOEFL test (adapted from Gilfert 1995)

	TOEIC TEST	**TOEFL TEST**
Listening Comprehension		
Part 1 – One picture, four spoken utterances	20 items	N/A
Part 2 – Spoken utterance, three spoken responses	30 items	N/A
Part 3 – Short conversation, four printed answers	30 items	25 items
Part 4 – Short talks, four printed questions and answers	20 items	25 items
Reading Comprehension		
Part 5 – Incomplete sentences	40 items	15 items
Part 6 – Error recognition – four underlined items per question	20 items	25 items
Part 7 – Reading comprehension - passages	40 items	30 items

Thus the near identical formatting and parallel test items leaves one asking how the TOEIC may claim to be a test of communicative abilities if the TOEFL cannot. Due to the paucity of TOEIC test research, this similarity is fortunate since it permits cautious extrapolation of TOEFL research findings to the TOEIC. This similarity is equally important when one considers the validity and reliability tests conducted, which will be discussed in Section 3.2.

As for exercising positive washback, many would question this since providing test-preparation has become a profitable industry (Hilke and Wadden 1997). While this issue will be discussed further in Section 3.5, it is worthwhile noting Robb and Ercanbrack's (1999) argument that if the format allows test-preparation then it is measuring test-taking skills rather than English abilities, an oft-made argument with M/C tests in general: though highly reliable, their validity is questionable.

3.2 Test Reliability and Validity

3.2.1 Reliability and the TOEIC Test

One prime concern of test users is test reliability. A test is reliable if multiple-administrations of an identical or near-identical test consistently yield highly similar results within one group (Bachman 1990; Bachman and Palmer 1996; Hughes 1989; Weir 1990).

As it is impossible to produce a test which is 100% reliable, the aim of test-writers is to produce a test that provides highly similar scores between administrations with the same examinees: "The more similar the scores….the more *reliable* the test…" (Hughes 1989:29).

According to Woodford (1982), group reliability of the TOEIC' listening test was

0.916 (a value of 1 being perfect reliability), with a standard error of measurement (SEM) of 25.95. This means that the scores may fluctuate +/- 26 points, or +/- 5 questions. For the reading test, reliability was found to be 0.930, with an SEM of 23.38; the total test had a reliability of 0.956 and the SEM was 34.93. Yet for individuals, scores obtained in both sections of the test are within 25 points of the candidates 'true score' only 67% of the time (TOEIC Technical Manual); 99% confidence in the score requires minimum gains of 63.5 points (*ibid*), thus weakening the test's reliability for individual scores. The standard error of difference (SED) was also calculated in the validity test to obtain measures of true gains for examinees. A 95% confidence rating that true gains were found requires minimum score increases of 65 points.

Referring back to Section 2.3, significant score-gains require 100 hours of language instruction. If these gains are interpreted as obtaining a 'true' increase, then the college's expectations of 80 point-gains in a 48-hour term are significant, indeed. Since many students do meet this goal, are the gains valid?

3.2.2 Test Validity and the TOEIC Test

Test validity can be evaluated from four different but related perspectives: construct validity, face validity, content validity and criterion-related validity (Bachman 1990; Hughes 1989; Weir 1990). These aspects can be defined as follows:

1. *Construct validity* is concerned with whether or not the test is actually testing the criteria it claims to test.
2. *Content validity* is concerned with the appropriateness of the tasks to test the desired criteria. In other words, do the tasks represent the target language use (TLU) and target language domain (TLD)?

3. *Face validity* is the degree of acceptability the test possesses in the eyes of the administrators and the testees and, as such, is a qualitative evaluation.

4. *Criterion-related validity* consists of both concurrent and predictive validity. If the results from one test format agree with results on a different test format, they are said to have concurrent validity. The predictive validity of a test is the degree to which the test accurately and consistently predicts the testees' future performance or behaviour.

The first administration of the TOEIC involved 2,710 Japanese participants. In conducting its validity study on both the direct and indirect measures of the test, ETS selected 500 examinees from this group based on their TOEIC score and placed them in groups of 100. Each group had scores approximating the following score-bands: 950, 765, 580, 315 and 45. Although all 500 people took the TOEFL to establish concurrent validity, only 20 examinees from each group of 100 participated in the direct measures validity tests (Woodford 1982), for a sample of just 4% of the total number of examinees.

Table 3.3 indicates that although the input for the validity tests for reading and listening was in English, the response tasks were conducted in Japanese making interpretation of the results and thus the TOEIC' construct validity circumspect. They may indicate comprehension abilities; they do not indicate communicative abilities. That this approach was taken is perplexing, for as Woodford (1978:1; 1982:3) explains,

> "..if the test is supposed to indicate how well you speak Arabic and doesn't require you to speak Arabic, it is of doubtful validity. […] If a language test is … to measure whether a person can read Japanese or not, then the person who scores high on the test should be able to pick up the Japanese newspaper and tell us what the lead article says."

While he does not specify that these tasks should be accomplished in the target

language, one would expect so on a test of communicative abilities; otherwise, the test is measuring only the candidates' comprehension abilities.

Table 3.3 A summary of the direct measures used in the validity test (adapted from Woodford 1982:10)

SKILL DIRECTLY MEASURED	TASKS USED
Listening Comprehension	- 25 taped English stimuli: 15 short statements 10 dialogues - candidates asked 3 questions for each item - examiner asked the questions in Japanese - candidates encouraged to answer in Japanese
Reading Comprehension	- unspecified number of reading tasks - a total of 30 questions asked in relation to the reading tasks - questions orally posed in Japanese - examinees orally answered in Japanese
Writing	- 3 subtests 10 dehydrated sentences write a 25-40 word business letter according to a specified context translate 10 Japanese sentences into English
Speaking	- Language Proficiency Interview (LPI), as used by U.S. state and local government agencies

ETS was also to "develop a procedure for score interpretation that would allow score recipients…to see typical samples of examinees writing efforts" (Woodford 1978:2). Unfortunately, the samples provided are based on the initial validity study and examinees were only required to compose a 25 – 40 word business letter (Table 3.3). This raises questions for the validity test for writing due to the M/C format of the TOEIC' Reading and Structural Knowledge sub-tests.

The results from the validity tests' direct measures of listening and speaking were compared with the TOEIC listening score results and the reading and writing results were compared with the TOEIC reading scores. The results of the validity test are

summarized in Table 3.4. As can be seen, the two listening tests have a very high correlation of .90; a correlation of 1 would indicate that both tests measure the same constructs to the same level of accuracy. The other three test aspects have respectable but lower correlations ranging between .79 and .83, perhaps warranting further validation with more samples for such sweeping claims of validity to be made.

Table 3.4 A summary of the correlation between TOEIC test results and the validity test results (Adapted from Woodford 1982:12-15)

TASKS	CORRELATION
Validity test listening and TOEIC listening	0.90
Validity test speaking and TOEIC listening	0.83
Validity test reading and TOEIC reading	0.79
Validity test writing and TOEIC reading	0.83

A study conducted by Moritoshi (2003) into the TOEIC found that the combined lack of operational definitions for the test and an over-reliance on concurrent validity weakened its construct validity. Furthermore, Moritoshi notes that

> "three of the concurrent tests were unvalidated and all were scored subjectively…. and no 'negative evidence' is offered to show that the tests are *not* testing other, unrelated abilities…" (*ibid*:12)

concluding that the construct validity needs further substantiation.

It might further be noted that ETS established concurrent validity between the TOEIC and the TOEFL, which as was demonstrated in Section 3.1.3 are essentially the same test produced by the same company. This problem of 'the dog chasing its tail' is discussed by Bachman (1990), who concluded that independent evidence in the form of

construct validation is required if interpretations of concurrent validity are to be made. With the weak construct validity pointed out above, arguments for the validity of the test undeniably become circular.

While few disagree that a person with advanced English communicative abilities can obtain a high TOEIC score, interpreting this as indicative of the TOEIC measuring communicative ability has significant implications for the application of test score results and for the way English is approached by educators and learners alike.

Bachman (1990) touches on this while discussing the pit-falls of predictive validity. The fact that a test perhaps provides inaccurate information about a person's ability is not serious in an educational setting since the error may be easily rectified (*ibid*). However, because M/C tests do predict placement levels with relative accuracy, proponents of these tests use the phenomenon as "evidence that they measure the abilities that constitute the criterion" (*op cit* 1990:251*;* also, Schmitt 1999). With regards to the TOEIC, this means that because in research situations candidates' scores correlate with ETS established bands of communicative abilities, it is interpreted that the test is a valid measure of these abilities even though the test has not been demonstrated to measure such criterion.

Spolsky (1989) wrote that valid language test development requires a definition of 'what it means to know a language', for if you do not know what you are measuring, you cannot measure it. The TOEIC is considered a test of English communicative skills. To know whether or not it is measuring communication, we must make clear what communication and the ability to communicate entails.

3.3 Communication and Communicative Competence Defined

3.3.1 Communication

Communication is defined as

> "…[involving] a 'reduction of uncertainty' on behalf of the participants" (Palmer 1978 cited in Canale 1983:4).

> "…the exchange and negotiation of information between at least two individuals…(it) involves the continuous evaluation and negotiation of meaning on the part of the participants" (Canale, *op cit*).

This exchange may be realized through a conversation or discussion between participants, a lecture, or between a written text and the reader (Brazil 1992; Swain 2001). In the case of the TOEIC' items (Appendix 1; Table 3.1), there is no option for the examinee to 'reduce uncertainty' or 'negotiate meaning'. The examinee's role is non-participatory in the listening tests; there is a salient lack of negotiation with the reading tasks, as well.

Referring to Canale and Swain's model of communicative competence, Spolsky (1989) and Widdowson (1989) write that knowing the grammar of a language is irrelevant if the speaker is ignorant to the rules of use. Coulthard (1985) argues that correct interpretation is essential in any act of communication, whether it is the interpretation of the sender's message or the interpretation of the receiver's knowledge. In other words, communicative competence involves much more than knowledge of the language's structure.

3.3.2 Communicative Competence

While 'communicative competence' has become a buzzword (Canale1983), calls for the inclusion of it within the language classroom are not new.

Brumfit and Johnson (1979) note the 60's shift from the teaching of language use to language structure, using syllabi that effectively institutionalized methodologies where we "present a structure, drill it, practice in context…then move to the next structure" (*op cit* 1979*:1*). EFL/ESL educators took the view that language acquisition was akin to laying rail-tracks, with learner evaluation developing along the same approach.

The 70's saw linguists and socio-linguists pushing for language classrooms to center on teaching communicative competence. Widdowson (1979) argued that focusing on form over the potential that language has in use is irrelevant if the end result is knowledge that can neither be accessed nor applied in communication, an argument which has been repeated often since then (Canale 1983; Cook 1989; Ellis 1990; Morrow 1979; Sinclair and Renouf 1988; Spolsky 1985; Widdowson 1989).

The lack of a clear-cut approach to incorporating communicative competence in the classroom (Hadley 1998) and in testing (Bachman 2000) 30 years later underlines the difficulty in defining it. Candlin (foreword in Weir 1990) remarks that this difficulty has resulted in test-developers creating 'communicative' tests that manifest themselves in the *accoutrements* of traditional testing paradigms. The TOEIC would be a case-in-point.

Canale (1983:2) suggests why this is the case:

> "…one also finds confusion and lack of consideration of many of the basic concepts involved …in the area of communicative language pedagogy results… from failure to consider and develop an adequate theoretical framework."

His description of communication led him to develop just such a framework (Canale 1983), which has been widely adopted (Chan 1986; Coulthard 1988; Ellis 1990; Weir 1990; and Richards and Rogers 2001). According to this model (Table 3.5), communicative competence is composed of grammatical, socio-linguistic, discourse and strategic competences.

Table 3.5 Framework for communicative competence (Canale 1983:6)

TYPE OF COMPETENCE	DEFINITION
Grammatical Competence	refers to the extent that mastery of the language code has occurred, including vocabulary knowledge, word formation, syntax, pronunciation, spelling and linguistic semantics
Socio-linguistic Competence	refers to mastery of the socio-cultural rules of use and rules of discourse; " the extent to which utterances are produced and understood *appropriately*…depending on contextual factors" (*ibid:*7) for example, the status of participants, the purpose of the communication and the conventions associated with the context
Discourse Competence	refers to mastery of "how to combine grammatical forms and meanings to achieve a unified spoken or written text" (*ibid:*9) suitable to the genre; includes use of cohesion and coherence.
Strategic Competence	refers to mastery of verbal and non-verbal communication strategies we employ during breakdown in communication or when we lack any of the competences to communicate effectively; also used to enhance the effectiveness of communication

Canale (*op cit*) notes the misconception of the status of socio-linguistic competence in SLA education. Although often considered secondary in importance to grammatical competence, he argues that the appropriateness of utterances is as important as grammatical correctness and in agreement with previous arguments (Section 3.3.1), this competence is essential for interpreting utterances.

Similar arguments have been made for the importance of discourse competence (Widdowson 1975, cited in Canale 1983:10; Hymes 1979; Coulthard 1985). While utterances may be grammatically and socio-linguistically correct and thus

'communication' has occurred, they may not be suitable for the genre or simply not cohesive / coherent, as the following examples demonstrate:

Ex. 1 What did the rain do?
The crops were destroyed by the rain. (Widdowson 1975:25, *op cit*)
Ex. 2 Where is my typewriter?
Your typewriter is in the cupboard.

While the responses in both examples are not incorrect, *per se*, the first example's answer is not discursively cohesive. Typically, old information precedes new, where an earlier repetition of 'rain' would provide a 'physical' link between the question and answer. The response in Example 2 is correct but is not what a native speaker would answer with. Instead, a typical rejoinder would be 'It's in the cupboard' or simply 'In the cupboard'.

Conversely, imperfect grammar can create more successful communication or be more readily understood than an utterance that conforms to the structural rules of the language:

> "…what to grammar is imperfect…may be the artful accomplishment of a social act (Garfinkel 1970) or the patterned, spontaneous evidence of problem solving and conceptual thought (John 1967:5)." (Hymes 1979:8).

This distinction between structure and function supports criticisms against M/C tests such as the TOEIC and other traditional language tests. As Morrow (*op cit*) and Rea (1978, cited in Weir 1990:3) argue, these tests do not provide evidence that the candidate can apply their understanding of the language to situations they might encounter in real-life. Richards (1980, cited in Larsen-Freeman and Long 1991:42) and Wu (1998) concur, noting that while the examinee may be able to choose the correct answer on a M/C test, " 'it is feasible that the learner, given the opportunity, would reject

all of the proffered alternatives.' " (Richards *op cit).*

Arguments for the acknowledgement of and need for a focus on the functional and pragmatic aspects of language are not propositions for a departure from structural studies as such (Terrell 1977, cited in Canale 1983:17; Widdowson 1989). Instead, they are calls for recognition that communicative competence requires competence in the four aspects summarized in Table 3.6 and it is these competences that should be the focus of the language classroom and tests.

While not universally accepted (for example, Tarone 1988, cited in Brown 1994:33), the most widely acknowledged description of what communicative competence embodies expands on the above ideas. According to Bachman (1990), communicative competence consists of two distinct components, organizational and pragmatic. Organizational competence includes grammatical and textual competence, while pragmatic competence involves illocutionary and socio-linguistic competence, each encompassing finer aspects.

Grammatical competence includes competency in vocabulary, morphology, syntax, phonology and graphology; in other words, the formal, structural aspects of the language. Textual competence includes cohesion and rhetorical organization, or coherence. Illocutionary competence is based on speech acts and language functions: what we say and how we say it to get things done. Socio-linguistic competence includes aspects of language as they are used in context: sensitivities to dialects or varieties of speech, register and naturalness, and the 'ability to interpret cultural references and figures of speech' (Bachman 1990:97).

What this means for the language classroom has come to the forefront of research in the past decade.

3.4 Theories in Second Language Acquisition

For SLA to occur and to attain the level of communicative competence outlined above entails developing the four-skills of the target language: reading, listening, writing and speaking; and while accepting that learner motivation is an influential factor, as are individual learning styles, beliefs and cognitive processes, research indicates that there are certain external elements generally applicable to all learners in the acquisition process. By external elements, I mean elements the teacher can introduce or foster in the class to help acquisition. It also means a better understanding of this process. The aspects that I consider in this section are 'focus on form', 'vocabulary acquisition', 'reading', and 'listening'.

3.4.1 Focus on Form

According to Long and Robinson (1998), three current approaches exist in L2 education, which they term 'focus on form**S**', 'focus on meaning' and 'focus on form'. Focus on form**S** (for example, Grammar Translation, Audio-Lingual, Contrastive Analysis) is a synthetic approach to language learning – language is focused on in discrete elements, practiced in isolation, the order and content as decided by the teacher. As researchers argue, however, acquisition is not linear (Laufer and Paribakht 1998; Long and Robinson 1998; Morrow 1979; Rutherford 1987; Schmitt 1998; Schmitt 1999):

> "Instead of learning discrete lexical, grammatical or notional-functional items one at a time, research shows that both naturalistic and classroom learners rarely…exhibit sudden categorical acquisition of new forms or rules… learners typically pass through stages of nontargetlike use of target forms, as well as targetlike and nontargetlike use of nontargetlike forms…"
>
> (Long and Robinson 1998:16-17).

Focus on meaning (for example, Suggestopedia, The Natural Way, Communicative Language Teaching) came in reaction to this. Language teaching was approached naturally with structural aspects of the language learned implicitly through natural exposure to the language (Long and Robinson 1998). The results indicated, however, that high fluency was attained but fossilization occurred at lower levels of ability.

An analysis of this polar situation left Widdowson (1989:131) asking two important questions:

1. Is it possible, in principle, to have grammatical ability without pragmatic ability?
2. Can we have pragmatic ability…without grammatical knowledge or the ability to compose or decompose sentences with reference to it?

While both questions can be answered affirmatively, Widdowson notes that an imbalance of the two indicates a lack of competence since not knowing the parameters of use results in unnatural utterances, as does a lack of structure. Thus, the pendulum swings back but moderately so. We need to teach structure, the question being what grammar, when and how (Doughty and Williams 1998).

This attention to 'focus on form' (FonF) (Ellis *et al* 2001; Doughty and Williams 1998; Hayashi 1995; Long and Robinson 1998; Swain 1998; Swain 2001; White *et al* 1991; Williams 1999) is the result of research that demonstrated the shortcomings of purely content-based learning situations (Doughty and Williams 1998; White *et al* 1991). The three basic premises behind FonF (Long 1991, cited in Ellis *et al* 2001:282-3) are:

1. Learners learn grammar as a product of communication rather than as a process.
2. The lack of information processing capacity results in difficulty in learning and producing correct forms.
3. Focus on form during communication is beneficial.

Although Willis (1990) has effectively argued against the effectiveness of structure-based lessons in a synthetic atmosphere, proponents of FonF stress the need of incorporating it into meaning-centered instruction. Effectiveness has been found to increase if the focus is learner-initiated rather than instructor-initiated (Ellis *et al* 2001; Williams 1999). White *et al*'s (1991) research drew similar conclusions. Interestingly enough, Williams' (1999) own research indicates that learners tend to focus on lexis rather than on structure.

3.4.2 Vocabulary Acquisition

While how vocabulary is learned is still unclear, as with structure it needs focused study and reinforcement (Ellis 1988; Sinclair and Renouf 1988).

As researchers maintain (Ellis 1988; Sinclair and Renouf 1988; McCarthy 1990; Willis 1990), grammar and vocabulary are two views of the same phenomenon. Willis (*op cit*: 86) argues that lessons should focus on lexis to provide "the framework for the gradual acquisition of the grammar" because the functions and collocates of words contribute to discourse organisation, a view also held by Sinclair and Renouf (1988).

Aitchison (1994) and Nagy (1988) would seem to agree, suggesting that vocabulary may be learned in chunks and as collocates and that the associations made between items and their meanings and uses during acquisition depends on the surrounding context. Research by Hulstijin and Laufer (2001) suggests that even with context the learners' retention is less if they are supplied with the definitions but are not required to use the items in a specific context.

With regards to vocabulary in testing, a clear definition of what lexical knowledge is determines how and what one tests – sentiments which echo Spolsky (Section 3.2.2).

As Laufer and Paribakht (1998:366) write,

> "…if lexical knowledge means the ability to use words in well-written sentences or discourse, it would then make little sense to measure lexical knowledge by …a multiple-choice test".

Unfortunately, it would seem that such a definition is non-existent. Schmitt posits that our lack of knowledge regarding vocabulary acquisition stems from previous research focusing on breadth of knowledge as opposed to depth (Schmitt 1998). By depth, researchers mean the levels involved in lexical competence: its form in both spoken and written texts, its grammar and the collocations it takes, its limitations as to how often and how it is used and its associations (Nation 1990, cited in Laufer and Paribakht 1998 and in Schmitt 1999).

This concern with only breadth has ramifications for testing, for example the TOEFL. Schmitt feels its imbedded in-text vocabulary test has "good technical characteristics, but the question about what it really measures still remains" (Schmitt 1999:190) because it lacks construct validity. He extends this claim to most vocabulary items and tests such as the TOEIC, concluding that future vocabulary tests must be developed in accordance with this concept of lexical competence. The same could be said for listening.

3.4.3 Listening in Testing

Researchers tend to draw comparisons between language listening skills and computer processing, possibly because both happen in 'real' time. In reading a text or in conversation, one may review previously encountered information or can ask for clarification. When one is passively listening, the information must be decoded and

understood as it is heard. Thus Wu (1998:23, citing Just and Carpenter 1992) rightly argues listening comprehension tasks have high demands: the storage and processing of information overloads the brain, and not everything can be remembered, perhaps explaining why learners appear to understand everything yet know nothing. In a listening comprehension test,

> "…[memory resources] is taken up when the testees struggle with language difficulties. The result is slowing down the processing …or simply leaves no room for processing meaning." (s*ic)* (Wu 1998:23).

This reasoning could be used in justification of M/C formats, such as the TOEIC, to test listening comprehension as they may reduce the task demands. However, Wu contends that M/C tests favour advanced level listeners but not less skilled listeners. Advanced listeners know what to discriminate for based on the question and options provided. For less-enabled listeners, syntax rather than contextual clues becomes their guide (Conrad 1985, cited in Wu 1998:27).

Wu's research also indicates that this format allows for uninformed guessing: participants were able to guess the correct options but for the wrong reasons. Conversely, participants would choose an incorrect option even though the information they had extracted was accurate. Thus for all levels of listeners, misinterpretation of options and unknown words make tests such as the TOEIC measures of reading comprehension and vocabulary, and not purely listening. This increases skepticism of the construct validity of M/C tests for listening, an argument also made by Weir (1990) and echoes that made by Schmitt (Section 3.4.2).

Some researchers and educators argue that SLA conceptualization affects testing practices, as does the converse. In other words, washback exists.

3.5 Washback in Testing

Washback is a phenomenon in which how we test is perceived to influence classroom practices, and syllabus and curriculum planning (Alderson and Wall 1993). Thus, if we believe that a discrete item, M/C test measures communicative abilities, this is how the language may be taught – as discrete items, which can be built on one another much like we lay rail-tracks (Brown in Leonard 1998; Hamp-Lyons 1998; Rutherford 1987).

As Leonard (1998) and Gorsuch (2000) write, the format of Japanese University entrance-exams runs counter to the injunctions of *Mombusho* to develop communicative abilities. These exams are still mainly M/C in format, test vocabulary and require translation. Tasks that test writing and aural/oral abilities are rare. Thus, students see no point in focusing on these skills at school and as a result the teachers ignore them.

Gates (1995) cites seven contributing factors to washback: face validity, accuracy, transparency, utility, monopoly, anxiety and practicality. With regards to the TOEIC in Japan, it has high face validity, is perceived to be both reliable and valid, and is thus transparent: the test meets the users' language needs. It is easy to administer and scores are reported back within one week. The test format may also have minimal effect on the anxiety levels of the test-takers since M/C tests are common to high-school tests and university entrance exams.

Childs (1995) and Hilke and Wadden (1997) state that contrary to ETS claims of the ineffectiveness of test-preparation, it does result in score gains, even if only in the short-term. As prospective employers often consider TOEIC scores, students' goals are to increase them. With a shrinking student population, university administrators develop language programs to accommodate this aim in trying to maintain a healthy student

population. At the institute focused on in this research, the language program's main goal is to raise TOEIC scores (Cunningham 2003). While there are many ways to approach this, often what happens is students focus on the skills necessary to increase their score: discrete grammar/vocabulary study and practice tests. While oral communication is a part of the program, it is held in lower regard to the test. Thus with regards to the TOEIC, washback is quite real. Of course, teachers may opt not to focus on test preparation (Smith 1991, cited in Alderson and Wall 1993:123; Wesdorp 1982, cited in Alderson and Wall *op cit:124*). However, Schneider states that,

> "…administrators have been unwavering in one point: the use of TOEFL [and TOEIC] gain scores as a key element in program goal setting and evaluation……..it can be said there was pressure … from administrators to get scores moving up". (Schneider 2001:35-36).

If common to most foreign institutions, this effect of washback has serious implications for language study. As was shown earlier (Section 3.1.3), one intention of the TOEIC was to positively effect language study. However, the type of washback currently felt could be described as negative. This focus on TOEIC preparation may not be as detrimental to SLA in ESL environments due to learners theoretically having daily opportunities to use the TL. Yet within EFL environments, the effect on communicative competence may be more acute.

To explore the relation between TOEIC test score gains and increased communicative competence, I developed the Test of Interactive Communication – TIC.

Chapter 4 TIC TEST METHODOLOGY

4.1 Test Purpose

Proponents and critics alike have voiced their support for or against the verity of the TOEIC test to indicate true communicative ability and the general application of the results of individual scores; however, very little independent research has been conducted into these essential questions:

1. Does the TOEIC test an examinee's ability to use the English language?

2. Are TOEIC test score gains valid indicators of increased NNS' English communicative abilities?

3. While NNS' with advanced communicative abilities may obtain high scores on the TOEIC, do those who obtain high scores have correspondingly high communication abilities?

4. Should low TOEIC scores be equated with low communicative abilities?

For this investigation, I created a Test of Interactive Communication (TIC)(Appendix 2) with the purpose of attempting to quantifiably measure a particular group of EFL students' aural and reading abilities and their abilities to communicate in written English information extracted from various texts. Gains or losses on TIC were correlated with the participants' TOEIC test results to determine if similar gains/losses were obtained on both tests.

4.2 Test Description

TIC consists of both a listening and a reading section, each one composed of five tasks (Table 4.1, overleaf). The tasks vary, with participants completing a chart or dialogue or summarizing the content of a text. Both input and expected output is in English and note taking is encouraged. The general content conforms to that of the TOEIC.

The tasks were chosen because they require the examinees to extract information from texts as they might in real-life: taking telephone messages, taking notes, providing summaries and interpreting charts. They are also required to demonstrate an ability to interact appropriately, though in a restricted sense, within three specified contexts: having a formal telephone conversation; having an informal conversation during a chance meeting; and writing a formal letter of complaint.

4.3 Test Trial

TIC was piloted in the Autumn 2001 term with first-year regular-class level students and students who were either in the honours level or were taking English electives. Generally speaking, the trial group's English ability was higher than the participants' abilities in this study.

Based on the pilot results, I replaced tasks that would be too difficult and made adjustments as to the speed of the recording and to the presentation and explicitness of the directions. As a result, the test administered in the spring term was tailored for the groups' level, while maintaining authenticity of content and task demands.

Table 4.1 Description of the tasks used in TIC

Test Task	Theoretical Construct	Operational Construct	Time
Listening Task 1 (TLU domain: real life; an aural/written task)	- the ability to understand and convey a recorded telephone message	- the ability to understand information in a recorded telephone message and record it onto a telephone message form	Script time: 25 seconds Writing time: 10 seconds
Listening Task 2 (as listening task 1)	- the ability to discriminate between numbers with similar phonology	- the ability to understand costs and shipping times for various mail options, and record them on a chart	Script time: 40 seconds Writing time: 10 seconds
Listening Task 3 (as listening task 1)	- the ability to understand spoken directions while looking at a map	- the ability to understand spoken directions while looking at a map; to trace the route described and locate the destination	Script time: 52 seconds Writing time: 10 seconds
Listening Task 4 (TLU domain: real-life/classroom - an aural/written task)	- the ability to take notes while listening to a short talk and to summarize the talk, in writing	- the ability to take notes on a weather report and to write a summary of the forecast	Script time: 50 seconds Writing time: 2 minutes
Listening Task 5 (as listening task 4)	- the ability to take notes on a short talk and to summarize the talk in writing	- the ability to take notes on the description of a modern circus, and to write a coherent summary of it.	Script time: 1 min. 52 sec. Writing time: 3 minutes
Reading Task 1 (TLU domain: classroom; reading/writing task)	- to demonstrate knowledge required to function in formal oral interaction; to demonstrate the ability to recognize discourse cues and cohesive elements, requiring certain responses; to demonstrate pragmatic knowledge related to manipulative functions and socio-linguistic knowledge	- to understand and be able to complete a written dialogue between two business acquaintances, which conforms to this genre of discourse;	Reading/writing time to fit in the allotted 45 minutes for the entire reading section
Reading Task 2 (as reading task 1, but within a social context)	As reading task 1, but for an informal, social discourse	As reading task 1, but for an informal, social interaction	As for reading task 1
Reading Task 3 (TLU domain: real-life writing)	- to demonstrate ability and knowledge necessary to write a formal letter	- to use key information to write a letter of complaint which conforms to accepted norms for this genre of writing	As for reading task 1
Reading Task 4 (TLU domain: real-life/classroom reading/writing task	- to be able to read and understand a short text with an accompanying chart	- to read a short text accompanied by a chart; to demonstrate an ability of how to use and interpret the chart by following written directions and answering 2 written questions	As for reading task 1
Reading Task 5 (as reading task 4)	- to demonstrate the ability to read a short text and the ability to write a coherent and cohesive text	- to read a short biography and summarize it in a written text	As for reading task 1

Adjustments were also made to the marking scheme. There originally were a limited number of points that could be awarded for comprehension, fluency, cohesion and errors. However, I felt that this did not adequately discriminate between various performances and therefore the scheme was set at "unlimited" where possible. Table 4.2 in Section 4.4 exemplifies this further.

4.4 Test-Task Construct

TIC measures comprehension abilities, fluency, vocabulary, syntax, cohesion, coherence and register. Points were calculated in single-unit increments and possible total scores were unlimited. Examples for each and the roles they have in the marking scheme can be found in Table 4.2.

As McCarthy (1990:12) writes,

> "…the division between grammar and lexis is not so sharp…any word in the language can be examined from the point of view of grammar and, vice versa, any word, even words like articles and prepositions can be considered as a vocabulary item"

making it difficult to categorize errors. In an M/C test, responses are either correct or incorrect. While this makes for unambiguous, fast marking, understanding whether or not the error would hinder communication is not possible. According to Hughes and Lascaratou (1982), this seems to be an important factor in native speakers' appraisal of NNSs' English. For these reasons, I have sub-divided lexis and grammar into two groups: Vocabulary 1, Vocabulary 2, Grammar 1 and Grammar 2. This is also in an attempt to understand where improvement exists in the participants' competence.

Table 4.2 Marking constructs and their roles in the marking scheme

CONSTRUCT ITEM	EXAMPLE	ROLE
Comprehension	Recording key words and concepts, either in note-form or in completing the task	1 point awarded per item, *ad infinitum*
Fluency	The use of complete sentences and phrases to convey understood information; includes individual clauses found in complex sentences	1 point awarded per item, *ad infinitum*
Vocabulary 1: derivatives, inflections and collocations/idioms	Derivatives: adapt → adaptable → adaptor Inflections: walk → walks → walking small → smaller → smallest Collocations: **give a** hand/lecture/smile fall in love	1 point deducted per error, *ad infinitum*
Vocabulary 2: Incorrect lexical selection due to phonological confusion or meaning difficulties; incomplete phrases or sentences	Phonology: confusion/contusion complain/cream Meaning: boyfriend/lover Incomplete phrase: I have complaints *(about the)* bus service.	1 point deducted per error, *ad infinitum*
Grammar 1: Intra-language / developmental errors	I *(am)* catching my plane at 6:00p.m. I *(have*) never been to Disneyland.	1 point deducted per error, *ad infinitum*
Grammar 2: Errors in syntax/ incorrect choice of structure	x What do you like kind food? X Tomorrow I have going to America.	1 point deducted per error, *ad infinitum*
Cohesion: the formal links made by vocabulary to relate ideas and sentences to one another	Referents: this, that, the following 3, the former Semantic: my dog, my pet, Charlie, the crazy thing Repetition	1 point awarded per link used, *ad infinitum*
Coherence: is the sense found in the text; is not found in the grammar and lexis.	"What did you do yesterday? Yesterday there was a plane crash." (← is not readily coherent, though 'yesterday is acting as a formal link) "Where were you yesterday? I telephoned you." (← there are no overt cohesive links, yet the phrase is understandable in most situations)	Marked on a scale of 0-2
Register: is the suitable use vocabulary and grammar items for a given context	Formal language in a business situation is appropriate, but not usually at the family dinner table.	Marked on a scale of 0-2, but only in specified tasks

Errors in Vocabulary 1 (inflections, derivatives and collocations) are grouped separately from Vocabulary 2 (lexical mis-choice because of phonology or meaning) because I feel that the former do not impede comprehension of the message. Errors due to the latter may put a strain on comprehension, forcing the receiver to rely more on the surrounding context to infer meaning.

Syntax errors are likewise divided: errors classified as Grammar 1 (intra-lingual) do not tend to cloud the message. Furthermore, the speaker has demonstrated awareness of the appropriate structure but has not yet fully incorporated it into use – knowledge of but not competency in use, whereas Grammar 2 errors (word order, incorrect structural selection) may indicate deficiency in knowledge of the form of the language or in the meaning of the structure.

4.5 Scoring Rubric

As Table 4.2 showed, points are accrued in increments of 1 point per correct or appropriate use and deducted in increments of 1 point per lexical or grammatical errors. However, not all test-tasks are marked for each construct (Table 4.3, overleaf).

With regards to Reading Task 5, it appears that many Japanese students have difficulty in summarizing texts according to western norms (McGregor 2002; McMurray 2002), tending to 'plagiarise' the relevant material. Thus, if the examinee has picked and copied pertinent sentences to summarize the text, then it is possible to get points for comprehension, fluency and coherence.

Conversely, if they selected and copied one paragraph, they can only receive credit for comprehension if the information copied is related to the main idea of the text; otherwise, they do not get credit for the response. This reduces the objectivity somewhat, but if they 'plagiaries' it is not possible to evaluate their use of grammar or lexis even though they may have indicated comprehension of the text. Others may simply copy the first paragraph without understanding or needing to process the information and thus their fluency becomes questionable, as well.

Table 4.3 Grading criterion applied to individual test-task items.

TEST-TASK ITEM	CRITERION APPLIED
Listening Task 1 – Telephone message	Comprehension – 1 point awarded per correct item recorded
Listening Task 2 – Post office	Comprehension – 1 point awarded per correct item recorded; maximum 6 points possible
Listening Task 3 – Map directions	Comprehension – Maximum of 2 points possible – correct route and correct location
Listening Task 4 – Weather Report	Comprehension – 1 point per correct item recorded or inferred; unlimited points possible Fluency – 1 point awarded per phrase or clause within a sentence Vocabulary 1 and 2 – 1 point deducted per error Grammar 1 and 2 – 1 point deducted per error Cohesion – 1 point awarded per lexical link Coherence – scaled 0-2 points
Listening Task 5 – The Cirque de Soleil	As for Listening Task 4
Reading Task 1 – Business Telephone Conversation	Comprehension – maximum 10 points Fluency Vocabulary 1 and 2 as Listening Task 4 Grammar 1 and 2 Register – scaled 0-2 points
Reading Task 2 – Conversation between friends	Comprehension – maximum 8 points Fluency Vocabulary 1 and 2 as Reading Task 1 Grammar 1 and 2 Register
Reading Task 3 – Letter of Complaint	Fluency Vocabulary 1 and 2 Grammar 1 and 2 as Listening Task 4 Cohesion Coherence Register – maximum 4 points
Reading Task 4 – BMI Chart	Comprehension – maximum 2 points
Reading Task 5 – Terry Fox	As for Listening Task 4

The above qualifications are relevant to the previously presented arguments against objective measures such as the TOEIC. The examinee may not understand the text or even need to process it, yet still be able to select the correct answer. TIC' element of subjectivity allows the marker to better understand the strengths and weaknesses of the students, a difficulty with M/C tests.

4.6 Participants

The test was administered to four groups of approximately twelve students, for a total sample of 50 students. All were freshmen, aged 18 years of age. The majority of the participants were male (35) with only 15 females. While all had the same teacher for speaking, the groups had different teachers for the TOEIC test-preparation and business-English classes (Table 4.4). Group A belonged to teacher X's class, Groups B and C belonged to Y's class and Group D belonged to Z's class.

Table 4.4 The four groups of participants and their TOEIC test scores

GROUP/TEACHER	NUMBER OF STUDENTS	TOEIC SCORE	REQUIRED GAINS
Group A / Teacher X	11 students	165 - 235	125 points average
Group B / Teacher Y	12 students	265 - 285	50 points average
Group C / Teacher Y	13 students	240 - 255	75 points average
Group D / Teacher Z	14 students	290 - 300	35 points average

The participants had entry TOEIC scores ranging from 165 to 300 with a TOEIC score goal ranging between 350 and 400 points, thus requiring a minimum increase of 71.25 points on average. Group 'A' needed an average minimum increase of 125 points.

ETS has indicated that candidates on the lower end of the TOEIC scale may see increases more easily than those on the mid- to high- end of the TOEIC scale and this rationale is accepted by the college; however, it does appear that ETS is considering the difference between candidates with scores under 450 and those with scores over 500. Thus, the expected gain differential between Groups A and D would appear to put Group A at a disadvantage, creating a high-stakes test situation for them.

Even if the above interpretation is queried, there is the question of the

recommended hours for significant gains to be observed (Section 2.3). The entire semester provides less than half of it, with 48 hours of instruction. Of this time, 24 hours are dedicated to TOEIC preparation and general business-language studies with the other time spent on oral communication. The oral class's required textbook is notion-functional, dealing with situations such as describing people, giving directions, telephone calls and ordering in a restaurant. The time allotted for each topic has rendered each one to the status of discrete point learning and hence their applicability for either TOEIC preparation or the facilitation of discourse competence is questionable.

The students did not volunteer for this experiment, thus their motivation and attitude towards TIC is an uncontrollable variable in the results. While their views of the TOEIC are unknown, it is assumed that they took it more seriously than they did TIC due to the TOEIC' high face validity and the high-stakes nature of it with regards to their passing the course and graduating.

4.6 Administration

The entry TOEIC test was administered on April 4th with classes beginning on April 12th. TIC was administered on the second day of their classes with the author: on April 16th for groups A and C, and on April 17th for groups B and D. Due to the small size of the groups, I was the only proctor.

Directions for the test were provided in Japanese so that students had a clear understanding of what the purpose of the test was and how to complete it. Directions for each task, including the allowed writing time for each, were also supplied in Japanese. Space was provided on the test paper for students to record their answers.

The listening test was pre-recorded (see Appendix 3 for the tape-script) to

guarantee equal quality of input for each group. Time for reading the instructions prior to each listening exercise, as well as time for writing their answers was included in the taped recording.

The second administration of the test was carried out in the same manner as the first, but two weeks prior to the exit TOEIC test. The tests from the first administration were marked by the author and by a second-rater who was located in a different prefecture and therefore unfamiliar with the students. There was no second-rater for the second administration. The results of the tests will be analysed and discussed in the following chapter.

Chapter 5 ANALYSIS OF THE RESULTS

This chapter proceeds by first describing the inter-rater results for TIC, then by comparing results between TIC and the TOEIC. A discussion of these results with regards to the research questions will follow, including the implications for the students in this particular situation, for learners in Japan and in other EFL settings.

Of the 50 participants, the results from 36 TIC test scores were used in the inter-rating. This decision was based on the content of the test results. If the participant's responses were solely in Japanese or if they had not completed the tasks requiring composition then they were excluded from the inter-rating. The results used in the TOEIC/TIC analysis are based on all 50 tests.

5.1 Inter-rater Reliability

Hughes (1989) writes that we cannot expect a perfect reliability co-efficient between raters on tests that require an element of subjectivity. M/C tests have only one correct answer and thus there should be no difference in the score obtained by different markers; however, when the test requires the markers to invoke a certain element of subjectivity the degree of accordance between them becomes weaker. Yet, Hughes is also very explicit: if the scoring is not reliable then the test results cannot be reliable.

Considering TIC does require some subjectivity (Section 4.5), the second rater was not experienced with the test and it was not possible to discuss various responses to the tasks, it is not surprising that the inter-rater reliability is not ideal.

Looking at these results, the total test score correlation using the Pearson Product Moment (PPM) equation is .6515, the totals for the listening tasks correlate at .4229, and the totals for the reading tasks correlate at .5657. These figures are less than robust and

the initial impression is that the test is weak.

However, the SEM and the standard deviation (SD) suggest that the raters were consistent in their marking. The difference in the SEM and SD between markers is narrow, the greatest difference between markers at less than 0.5 points.

Furthermore, an examination of scorer-reliability for each of the competences measured by TIC (Table 5.1) indicates a consistently stronger correlation between their marking. The values are also comparable to coefficients cited by ETS in its validity study (Section 3.2.2).

Table 5.1 Scorer reliability for the competences measured by TIC

Variable	Correlation Coefficient, N = 36
Total comprehension	r= 0.8554
Total errors (grammar and lexis)	r=0.8305
Total cohesion	r=0.8244
Total coherence	r=0.7093
Total fluency	r=0.7524

It should be noted that this consistency drops somewhat when these aspects are considered separately for listening and reading. Perhaps part of the reasoning behind the weaker correlations are attributable to the argument by Wu (1998)(cited in Section 3.4.3) and made by O'Malley *et al* (1989). Many students are able to take notes but due to overloading of their processing abilities are not able to parse the phrases into any meaningful text, resulting in incomplete tasks. Therefore, while their comprehension may be high they cannot communicate the information. Unfortunately, correlations depend on larger numbers of samples for reliable results: the accepted minimal norm

appears to be 30 samples (Owen *et al* 1997). With the samples in the inter-rater study totaling 36, incomplete tasks could have set the quantity below 30 for a given task.

The inability to complete the listening tasks might have affected the students' confidence and these feelings influenced their performance on the reading tasks. As a result, there may be too few complete scores for each section to obtain accurate inter-rater correlations for the separate competences.

It must be emphasized that this is preliminary research and further development of TIC and other tests must be conducted before any conclusive interpretations may be possible. Nevertheless, quantitative analysis may only provide information to a certain depth; a clear picture needs a combination of both quantitative and qualitative analysis, which this particular project may provide.

While the TOEIC test's format may allow for a wider sample of structures to be tested and it can be scored with a high degree of reliability, TIC provides multiple opportunities for the participants to demonstrate their ability to actually interact with the language. Furthermore, the sample size for the TOEIC/TIC test results' analysis totals 50 examinees; this is considered large enough to supply reliable coefficients when using the PPM equation, yet I feel is small enough to permit a qualitative examination and discussion of the results.

Although the inferences drawn from an analysis of the separate competences must be cautious, it is the author's contention that the overall inter-rater correlations are sufficiently reliable to proceed with a comparison of results for both the TOEIC and TIC, allowing tentative conclusions to be drawn with regards to the research questions.

5.2 Analysis of the TOEIC and TIC Test Results

5.2.1 The April and July Test Score Results

Two questions posed in Section 4.1 concerned the interpretation of TOEIC scores:

1. While NNS' with advanced communicative abilities may obtain high scores on the TOEIC test, do those who obtain high scores have correspondingly high communicative abilities?

2. Should low TOEIC scores be equated with low communicative abilities?

These questions are addressed through a comparison of the results obtained on both the TOEIC test and TIC.

Referring back to Table 4.4 in Section 4.7, Group A had the lowest TOEIC scores and Group D had the highest. These scores were interpreted by the college as Group A having the weakest language skills within the elementary level classes and Group D having the strongest.

Table 5.2 indicates that the TIC score rank order of groups is somewhat different, as are the score variances. Group B, ranked 3rd in abilities by the TOEIC, ranks 1st with TIC. Furthermore, all four groups performed at similar levels on TIC, with scores differing by a maximum of 10 points.

While TIC exhibits a wider SD than the TOEIC does for total test scores, it must be remembered that the students were grouped according to their total TOEIC score in an attempt to create level homogeneity, with only Group A consisting of learners with a wide range of TOEIC scores. The SD on TIC for all groups is relatively similar, perhaps indicating that as a group they performed at comparable levels.

Table 5.2 Rank order of groups for the April TOEIC and TIC

TOEIC Test Group Averages				TIC Group Averages		
Total Test Score	Score	Rank	SD	Score	Rank	SD
A (N=11)	215.00	(4th)	17.89	37.86	(4th)	18.13
B (N=12)	249.58	(3rd)	5.82	48.08	(1st)	21.31
C (N=13)	280.00	(2nd)	5.40	40.49	(3rd)	19.94
D (N=14)	292.14	(1st)	3.78	42.10	(2nd)	13.25
Total Listening Score						
A	129.55	(4th)	16.65	11.23	(3rd)	6.71
B	142.92	(3rd)	23.30	14.21	(1st)	6.66
C	170.36	(2nd)	20.80	11.12	(4th)	6.80
D	170.38	(1st)	23.93	11.89	(2nd)	6.28
Total Reading Score						
A	85.45	(4th)	24.54	26.64	(4th)	13.84
B	106.67	(3rd)	23.39	33.88	(1st)	17.08
C	109.23	(2nd)	26.60	29.38	(3rd)	14.55
D	121.79	(1st)	20.53	30.21	(2nd)	9.18

Similar observations were made with the July test results (Table 5.3). As with the April tests, Group A ranked 4th and Group D ranked 1st in average TOEIC test scores, with Group B ranking 1st in all three categories for TIC.

Although the similarity between group SD and scores for TIC may suggest that all learners performed at comparable levels, the groups with the lower scores have a slightly wider SD range. Using the SD to create score bands, one can see that the extremes for each group overlap. Using the total reading scores for July as an example,

Group A's score-band is 23.43 to 53.99
Group B's score-band is 37.54 to 60.80
Group C's score-band is 22.66 to 68.04
Group D's score-band is 31.51 to 53.41

it would appear that the groups are not as homogenous as the TOEIC scores would indicate. Doing this exercise with all scores reveals similar results.

Table 5.3 Rank order of groups for the July TOEIC and TIC

TOEIC Test Group Averages				TIC Group Averages		
Total Test Score	Score	Rank	SD	Score	Rank	SD
A	281.36	(4th)	27.21	60.73	(4th)	23.13
B	302.50	(3rd)	51.37	74.13	(1st)	16.74
C	327.69	(2nd)	61.93	64.69	(2nd)	26.57
D	357.86	(1st)	64.08	62.79	(3rd)	13.48
Total Listening Score						
A	182.73	(4th)	36.49	21.67	(3rd)	8.67
B	185.42	(3rd)	31.29	21.96	(1st)	6.56
C	202.31	(2nd)	45.17	19.00	(4th)	8.32
D	215.00	(1st)	48.20	21.89	(2nd)	8.54
Total Reading Score						
A	98.64	(4th)	23.57	38.71	(4th)	15.28
B	116.67	(3rd)	29.26	49.17	(1st)	11.63
C	125.38	(2nd)	38.10	45.35	(2nd)	22.69
D	142.86	(1st)	33.61	42.46	(3rd)	10.95

A third observation is the noticeable difference between group scores on the TOEIC test and on TIC. Of course, scores increase in increments of 1 point with TIC and 5 points on the TOEIC. However, a simple calculation of dividing the TOEIC scores by 5 still indicates a larger difference between TOEIC group scores than those found with TIC.

Consequently, while the TOEIC test creates one picture of the groups' language proficiency TIC indicates a wholly different one when it comes to actually interpreting texts and using the language. It would appear that students are much closer in ability when it comes to language competence than the TOEIC test scores would demonstrate.

It also suggests that the TOEIC was not an accurate method for determining group levels for these learners.

Thus, the tentative answers to the two questions posed at the beginning of this section would be that low TOEIC scores should not be equated with low communicative abilities and those with higher TOEIC scores do not necessarily have higher communicative abilities. This last claim must be qualified since the groups in question, while having different score ranges which are interpreted as being equitable to different abilities by the university, are still recognized as being at the elementary level. The above findings also imply that the use of the TOEIC as a placement test should be reconsidered.

One salient implication of the above interpretation for these particular students is with their course grades and their four years of university study, especially for Group A. If it is accepted, then essentially all 50 participants have comparably similar language abilities. In fact, Group B appears to have stronger abilities than either Groups C or D, which have higher TOEIC scores. However, the TOEIC score gains that students in Groups A and B require are much higher and as was argued in Section 4.7, this may create additional pressure for them in passing the course. Furthermore, those students stand a higher chance of having to take supplementary English classes, creating a heavier course-load. As English is not offered as a major at this college but it is a graduation requirement, the potential negative repercussions are decreased motivation and increased antipathy towards the language.

Extrapolating these interpretations to a wider sphere, the use of the TOEIC test to decide who has the language abilities for specific work positions or gains entry to an academic institution perhaps requires serious reconsideration. The possibility exists that

a candidate who does poorly on the TOEIC may in fact have similar or higher communication skills than another who obtains a higher score, yet be denied access to employment or training opportunities.

While anecdotes exist with regards to how TOEFL [TOEIC] scores from some countries are treated differently than from others (Hamp-Lyons 1999), the establishment of score bands as criteria for university admission, language placement or hiring should be used in co-ordination with either a written or oral interview, even at the lower score levels. Interestingly, this process appeared to have been utilized prior to the TOEIC' adoption as an effective, all-purpose test of communication abilities (The Reporter #9); it is also akin to the approach taken by the *Eiken* test (Section 3.1.3).

Although the above analysis deals with how groups fared on both tests, indicating that high TOEIC scores may not necessarily be equated with high communicative abilities and *vice-versa,* it does not address the questions relating to TOEIC test-score gains correlating with improvements in actual language use.

5.2.2 TOEIC and TIC Test Score Correlations

One key research question concerned score gain correlations: Do gains in TOEIC test-scores reflect increases in communicative abilities? Total score gains, total listening and total reading gains for both tests were calculated to obtain a correlation, using PPM. The results can be seen in Table 5.4.

The gains for the listening sections have a strong, positive correlation. The significance means that there is a strong possibility (p=0.181 or 91%) that the TOEIC listening gains will increase in the same manner as gains on TIC. In contrast, gains for the reading tests have a low negative correlation indicating that as TOEIC reading scores

increase, TIC reading scores will decrease. However, with p=0.609, the possibility of this relation randomly occurring is 39%.

Table 5.4 Correlation between TOEIC and TIC score gains

	Coefficient	Significance
Total Listening Score Gains	+0.8193	p=0.181
Total Reading Score Gains	-0.3908	p=0.609
Total Test Score Gains	-0.9664	p=0.034

Referring back to Tables 5.2 and 5.3, all four groups had significantly higher TOEIC listening scores than reading scores yet higher TIC reading scores than listening scores. The negative correlation for reading would seem to demonstrate the difference in reading gains between TOEIC and TIC. As the listening scores are also inverse to the TOEIC scores, I would have expected a negative correlation with the listening tests as well.

However, recalling the arguments made in Sections 3.4.3 and 5.1, research suggests that learners can understand spoken information but have difficulty in processing it. With the TIC listening tasks, the students were able to write down key words they had heard but were not able to use the information to summarize the texts. This may explain the high correlation between the two tests' listening measures.

Perhaps TOEIC results are indicating the examinee's ability to extract key words from a spoken text and use them to guess the correct option. This supports the arguments that M/C tests are not measuring the examinee's ability to understand the content of a spoken text.

With regards to TOEIC test-takers in Japan, the balance between the test's

reading and listening sub-tests seems to be the observed norm: listening scores are generally higher than reading scores. Considering the approach to foreign language education with its emphasis on grammar-translation, the reverse would be expected: high reading scores but lower listening scores. As this expectation is borne out with TIC, it would appear that TIC test scores are creating a more realistic picture of the students' listening and reading abilities than the TOEIC test results would suggest.

Proceeding to the total test score gains, there is a near perfect negative correlation of –0.9664, meaning the more gains made on a TOEIC test the fewer gains made on TIC, and *vice versa*. This finding suggests that a student's ability to analyse a sentence for syntax and lexis and their ability to use appropriate grammar and vocabulary in a written text require distinct skills. It would also insinuate that there is little relation between reading or listening to a text and selecting an answer from a choice of four possibilities, and encountering a text, understanding the information within and being able to communicate that information in English. Both of these findings relate back to the SLA arguments made in Section 3.4.

This is a highly plausible interpretation where EFL learners are concerned, particularly in situations where studying for the test is emphasized over communicating in English (Robb and Ercanbrack 1999). However, this claim would demand further substantiation with learners of varying proficiency in varying situations for it to be generalised to all learning contexts.

To summarize thus far, the data obtained for this particular group of learners at this stage in their learning process suggests that TOEIC test score gains do not correlate with increased communicative competence. In other words, the gains made on the TOEIC are not necessarily valid and consequently for the learners in this study, it does

not appear to be an accurate measure of their abilities to use the English language.

This would support the conclusions drawn in Section 5.2.1 that the TOEIC should not be used to evaluate students and employees. The results may distort the picture of their true language abilities either in their favour or against them. This effect should concern employers as well, given that over-evaluating a worker's ability to communicate in English, either with interpreting incoming information or transmitting information to a foreign counter-part could have serious repercussions for the company.

The final question posed was regarding TOEIC test preparation and improvement in both fluency and accuracy. This question will be considered next.

5.2.3 Lexis and Syntax Accuracy Gains on TIC

This section deals with TIC results only since the TOEIC test results do not include a breakdown of the types of errors made.

As was summarized in Section 4.7, all students studied TOEIC preparation twice a week using a TOEIC reading-exercise practice textbook and a general business textbook. Each group had a different teacher and as a result the material covered varied between groups. Furthermore, it is unknown how these classes approached test preparation and how or what language competence development was included in the lessons. The other two classes per week were for conversation, using a notion-functional textbook. Very little explicit grammar teaching was conducted.

Table 5.5 indicates the results of the students' average gains on TIC with regards to comprehension, fluency, vocabulary, grammar, cohesion and coherence. Comprehension gains were impressive; it would appear the participants understood twice the amount of information during the second administration. These gains were

evenly distributed between the four groups, although Group B saw slightly larger increases.

Table 5.5 Average gains on the competences measured by TIC

Competence	April Test	July Test	Gains
Comprehension	79.3	158.86	+79.56
Fluency	99.05	130.44	+31.39
Vocabulary 1	-24.24	-24.60	-0.36
Vocabulary 2	-27.32	-33.89	-6.57
Grammar 1	-1.44	-3.25	-1.81
Grammar 2	-12.36	-13.72	-1.36
Cohesion	35.03	42.60	+7.57
Coherence	11.16	14.06	+2.90

An argument could be made that using the same test for both administrations contributed to these gains; the author would disagree, however, due to the test neither reflecting class content nor class activities, as well as because of the three-month time span between administrations.

Fluency would also seem to have improved, albeit moderately so. As would be expected with gains in fluency, there were increases in the errors made. What is interesting are the types of errors made. Lexical errors comprised 79% of the total errors made in the April tests and 78% of those counted in the July tests. While there was a balance of Vocabulary 1 and 2 errors (see Section 4.4) made in April, Vocabulary 2 errors increased in the 2nd administration's results; the percentage of Vocabulary 1 errors remained the same

Thus it appears that over the course of one term the students' fluency increased

somewhat, although their accuracy did not visibly improve. This would appear to support the observation made in the previous section with regards to the correlation between TOEIC and TIC test score gains: improving on a M/C test of grammar and vocabulary is not equivalent to improving one's accuracy in language use. The results also lend support to arguments against the effectiveness of discrete-item learning and the fallacy that language learners will progress in a linear fashion.

It must be kept in mind that a causal relation between the language program and increased fluency has not been established. The increase in fluency could be due to the program; it could also be attributed to lower anxiety levels within the students, since a one-hour long test on the second day of classes potentially caused them stress; furthermore, they had developed a relationship with their teacher and thus may have been more relaxed during the second administration. At the same time, it does not appear that test-preparation had any effect on the development of the learners' structural competence.

The students' use of cohesive elements slightly improved over the course of the term, although it appeared that the types of cohesion used did not differ between the two tests. Lexical repetition visibly outweighed the use of other cohesive elements, with the sporadic use of anaphoric and cataphoric noun phrases, such as 'this problem', and 'many problems'.

On both test administrations, coherence did not appear to be problematic for the students. Whether this is due to the tasks having established guidelines for how they could respond or to their EFL organizational skills is not clear. This would involve a separate research undertaking.

Similar findings were made with register: there were no noticeable differences

between test administrations. I feel that this stems largely from too few tasks being specifically designed to measure this particular competence.

While an in-depth discussion is not relevant to the main research questions, some implications of these findings for the language classroom are suggested in the following section.

5.3 Implications for the Second Language Classroom

Without conducting an analysis of classroom practices and content, it would be difficult to establish any sort of causality between the college's English program and the gains/losses in language competence measured by TIC. However, the findings will be used to hypothesise implications of washback from the TOEIC test via a test-preparation program.

The group of learners in this study demonstrated a stronger command of grammar than of vocabulary, which would be expected based on the *YAKUDOKU* tradition of language learning. As was described in Chapter 2, Japanese students tend to have high structural knowledge of English but weak pragmatic abilities. Whether this weakness stems from a lack of vocabulary depth or simple inexperience in using the language to communicate is unknown. However it raises questions as to the propensity to focus on syntax as opposed to vocabulary or discourse development.

As claimed in the preceding section, the TOEIC test-preparation program did not appear to result in improved structural competence. Although not significant, the numbers suggest the students held their ground, which supports arguments against the time-efficiency of allocating so much time to one language aspect if it does not result in noticeable change (Willis 1990). Thus for Japanese learners, and one would suspect for

learners in countries with similar approaches to EFL education, explicit grammar instruction should perhaps be relegated to a secondary position with more emphasis placed on other language competences.

Referring back to Section 3.4, advocates of focus on form (FonF) posit that learning occurs because of communication, focusing on form during communication aids the acquisition process and the inability to process information results in difficulties in learning and using correct language. It was also suggested that learners tend to focus on vocabulary related difficulties over grammatical ones.

Whether due to the course content or to the students developing confidence in their abilities is not clear, but it appears that the language program was conducive to them further developing their comprehension abilities and their fluency. Therefore, instead of focusing on test-preparation, it would be fruitful to adopt a FonF approach to see if exploiting the confidence factor results in improved structural competence.

Exposing students to texts that contextualise lexis could lead to development in depth of vocabulary, resulting in better reading comprehension. Learners also need to increase their processing capacities, especially with regards to listening. Grounding the vocabulary and listening aspects in communication-oriented situations as proposed by FonF supporters might further the development of all aspects of both written and spoken competences.

As Hilke and Wadden (1999) state, obtaining a high test-score is the goal of some learners, in which case test-preparation should be offered. However, the test and the preparation course should not be marketed under the guise of either improving their language abilities or as a measure of communicative abilities. At the same time, this singular goal of test-score gains highlights the negative washback of the TOEIC: the

score has a higher value than the skill it is supposed to be measuring.

As was shown in Section 3.1.2, the resulting negative washback from not questioning the claims of ETS and its affiliates is realized with test-users equating the TOEIC score with communicative abilities, as well as reinforcing discrete-item learning as an effective method in developing L2 communicative abilities.

Chapter 6 CONCLUSION

The goal of most language learners is improved written and/or oral communication skills. Unfortunately, measuring the competences involved is complex, time-consuming and costly. Thus with its relative low-cost, high reliability and quick marking turn-around, the multiple-choice TOEIC test has been adopted worldwide as a measure of EFL/ESL learners' language abilities. Unfortunately, little independent research has been conducted into the question of whether or not the TOEIC does, in fact, measure communicative abilities.

To address the question, the author designed and administered a direct measure of listening, reading and writing abilities (TIC) to a homogenous group of EFL learners in a Japanese college. The students took an entry and exit TOEIC test, with administrations of TIC paralleling the TOEIC tests. Test results were statistically and qualitatively analysed and compared, with the results suggesting that the TOEIC does not measure communicative competence.

To briefly summarize the findings from Chapter 5, it would appear that:

1. based on a comparison of the rank order of group scores for both tests, high TOEIC scores do not imply high communicative competence and low TOEIC scores do not imply low abilities, within this group of learners

2. there is no positive correlation between TOEIC score gains and increased communicative competence, as measured by TIC

3. a TOEIC test-preparation course does not necessarily result in improved structural competence during language use

4. the more predominant type of errors made by the students are lexical rather than syntactic

There are three salient implications of this. Firstly, without the inclusion of a more direct measure of communicative abilities, the TOEIC should not be used as a high-stakes test to evaluate an EFL learner, as this particular study has not demonstrated it to be a valid indicator of communicative competence.

Secondly, the TOEIC should not be used as an achievement test since it neither reflects class content nor what students are able to do with the language. This raises questions as to whether the TOEIC should even be used as a general proficiency test if *proficiency* is interpreted as the ability to do something, in this case use the English language.

Thirdly, it would seem that more effort should be expended in helping learners to develop the lexical aspects of their structural competence rather than the grammatical since vocabulary appears to cause learners more problems.

Being a preliminary study with only a small sample of learners, the conclusions drawn cannot be definitive. There is ample room for future research with regards to TIC, testing communicative competence and the TOEIC as such a measure.

This should take the form of additional TIC reliability and validity establishment, not only within a homogenous group such as this study involved but also with groups of varying language abilities in different learning environments.

Moreover, TIC is concerned with listening, reading and writing abilities. Speaking, on the other hand, is different from written communication because it happens in the moment. Unlike written communication, it is difficult to revise the spoken word. Thus independent research needs to be conducted into the relationships between TOEIC test-preparation and spoken communicative competence.

Other possible venues for related study include an analysis of how different

approaches to test-preparation affect TOEIC score gains, again in relation to communicative competence.

Finally, further work needs to be done on developing reliable qualitative tests that are easy to mark. It is difficult to make direct tests truly quantitative as decisions regarding what is correct or incorrect and what hinders communication and what does not are required at the most basic level, involving differing viewpoints as to what is grammar-related and what is lexis-related.

To achieve a direct test of communication that is highly reliable, allows valid interpretations of communicative abilities and has the administrative benefits of M/C tests in general will require substantial time and effort. Prior to this, educators need to define and accept definitions of what communication and communicative competence entail and these definitions need to be used as the foundation for improving classroom practices.

It is acknowledged that some test-takers are more concerned with simply improving their TOEIC score; it is also accepted that studying for the TOEIC is motivational for many learners. At the same time, all test users need to be fully aware of the intended purpose of the TOEIC and its limitations, and need to question many of the claims made by the developers and administrators of the TOEIC test. Most people involved recognize the high face validity associated with the TOEIC; to have so much riding on one test, yet not understand just what the test measures is to place too much faith in a company that profits from this misconception.

Understandably, such decisions require time to read and sift through the TOEIC material. However, if a test that is easy to administer, highly reliable in its marking and a valid indicator of language ability is desired, the time spent on understanding the test is

time well-invested, both for those interested in streamlining decision processes and for those with their future riding on the test score. Otherwise, the reasons provided for using the test become questionable: is the test used for how it is held to evaluate language learners' communicative skills, or is it adopted for the ease and convenience it provides educators and program administrators?

In conclusion, the author feels that the findings of TIC suggest the TOEIC should not be used to determine communicative abilities and it should not be sold as a measure of these abilities. To continue to do so will maintain the negative impact on SLA. Although the findings are suggestive, the author hopes they will act as an impetus for comprehensive research into the questions of the TOEIC test as a measure of communicative abilities and of how to test communicative competence.

Appendix 1 – Example of Test Items for the TOEIC Test

(Out of respect for copyright, this has been removed from this version of the dissertation).

Appendix 2 – The Test of Interactive Communication(TIC)

(Note: the test has been condensed in size for the purpose of space conservation)

PART 1 – LISTENING TEST

Activity 1: これから留守番電話のメッセージを聞きます。オフィスでは Michelle, Mona, Jake が一緒に働いています。注意深く聞いてテレホンメッセージの用紙を完成させて下さい。 書く時間は１０秒です。

Date : _____ Time: _____

To : _____

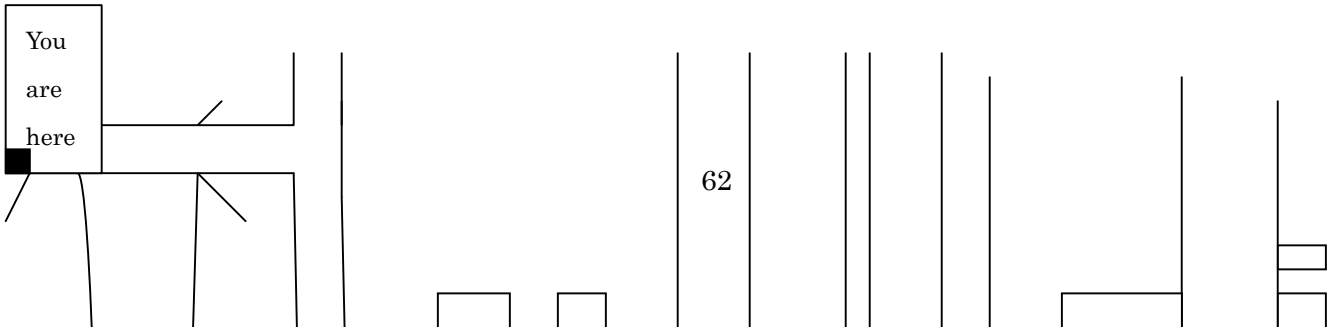
From : _____

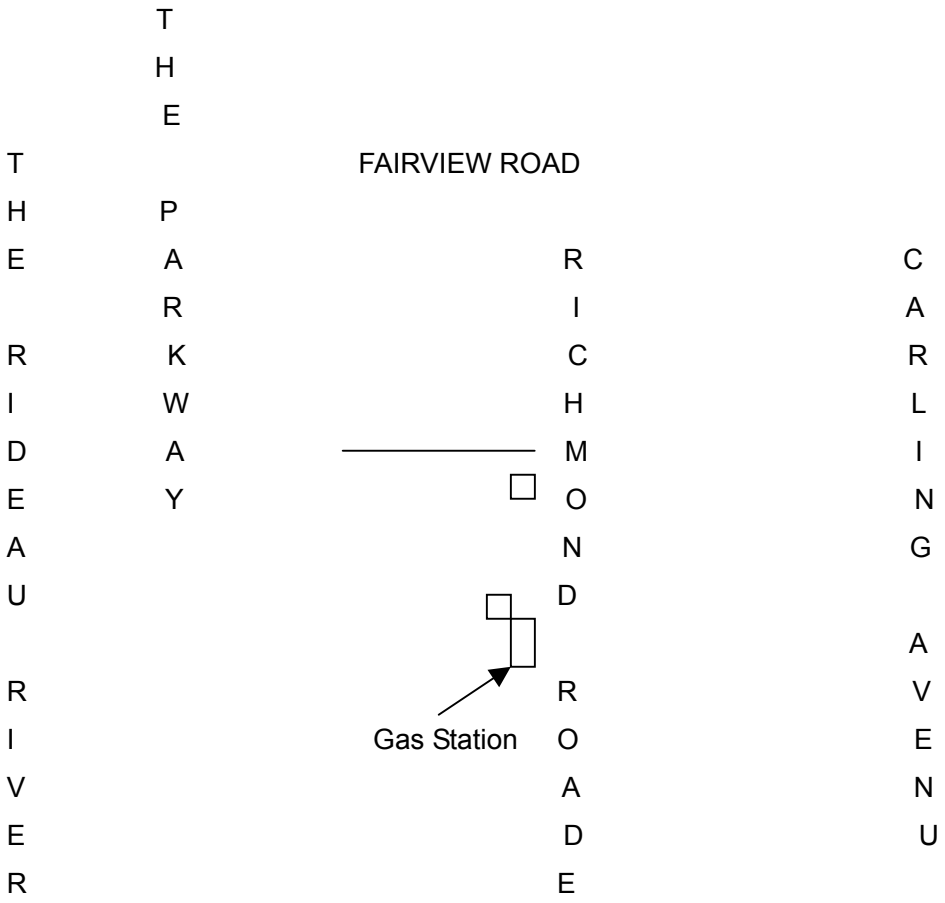
Message:

Activity 2 : アメリカへ小包を送るために郵便局へ来ている人がいます。その人に郵便局長はいくつかの違った郵送方法とそれぞれの料金を説明します。それぞれの料金と着くまでどれくらい時間がかかるかを書きとめて下さい。書く時間は 10 秒です。

	COST	SHIPPING TIME
- airmail	_____	_____
- surface	_____	_____
- SAL	_____	_____

Activity 3 : 次のページの地図を見てください。これから、あなたの友だちが家まで行く方法を説明します。説明を聞きながら 地図を追って下さい。友だちはどの建物に住んでいますか？ 書く時間は 10 秒です。)





Activity 4 : これから 天気予報を聞きます。 メモを取りながら 手短かにどんな天候だった

かを要約して下さい。書く時間は 2 分です。

_____

_____

_____

_____

_____

Activity 5 :　これから　サーカスについての短い講義を聞きます。メモを取り、それから　友だちに説明するように要約して下さい。書く時間は３分です。

_____

_____

_____

_____

_____

PART 2 – READING AND WRITING TEST　こ章からは４５分あります。

Activity 1: 次の会話を完成させて下さい。　これは二人のビジネスパーソンの電話上での会話です。

Mr. Mohamed Ali: Good afternoon, Mohamed Ali speaking.

Mr. Frank Pauls : _____.

Mr. Ali : I'm very well, thank you. And yourself?

Mr. Pauls : _____. Mr. Ali, I'm calling to see if it's possible to re-schedule our meeting on the 12th for another day. I've been called away on an urgent business trip.

Mr. Ali : _____

Mr. Pauls : Let's see… the 18th fits in with my schedule.

Mr. Ali : _____

Mr. Pauls : Yes, I'm available on the 19th, as well. What time suits you?

Mr. Ali :_____

Mr. Pauls : That's fine with me. So, I'll see you on the 19th at 9:30 a.m., then.

Mr. Ali : _____

Mr. Pauls : Good bye.

Activity 2 : 次の会話を完成させて下さい。これはしばらく会てない友だちどうしの会話です。

John : Mary! It's good to see you again. How have you been?

Mary : _____

John : Pretty good, thanks. Shannon and the boys are fine. In fact, we just got back from the cottage on Sunday.

Mary : _____

John : Are you familiar with the Westpark area?

Mary : _____

John : Well, it's on the north side of the lake, near Jackson.

Mary : Oh, that's a lovely area! I went driving through there once.

John : What about yourself? What have you been up to?

Mary : _____

_____

John : Well, that's great!

Mary : Yes, I'm happy. You know, I've got to run, but let's get together real soon.

John : O.k.! I'll give you a call later this month. Take care.

Mary : _____

Activity 3 : 下の箱の中にあるインフォメーションを見てください。

.

- fares too expensive
- buses → too few routes
  - → poor schedule
  - → no late night service
  - → seldom on time
  - → not enough seats

あなたはひんぱんにバスを利用していますがバスサービスに不満があります。 箱の中のインフォメーションを使って市の交通局あてに不満の手紙を書いて下さい。

_____

_____

_____

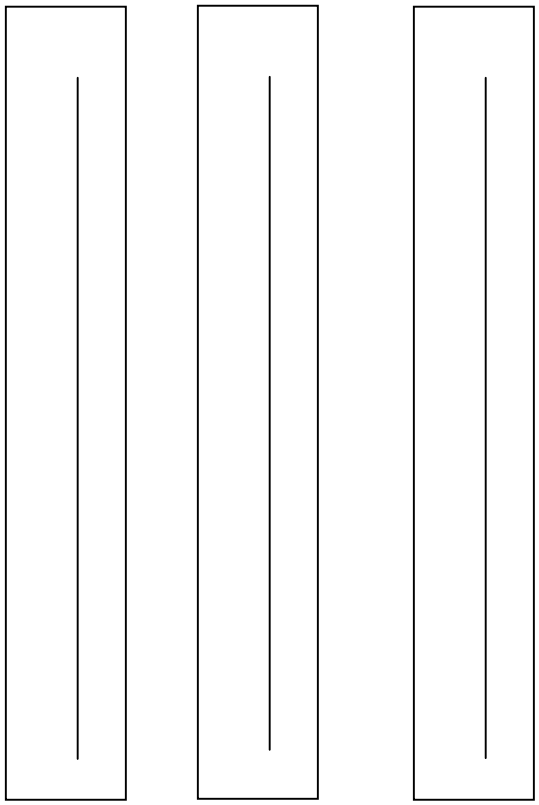
_____

_____

_____

Activity 4 : 私たちは常に体重と健康について考えます。私たちの体重が 正常値かどうか見分

ける方法として BMI チャートを使用する事があります。下の箱中のインフォメーションを見て質問に答えてください。(excerpt from Maclean's Magazine, 1/11/1999 pg.57).



The Body Mass Index uses height and weight to determine how close a person is to a desirable weight. Extend a straight line from your height through your weight and into the BMI column. The results:

30 AND OVER

You may be at risk of serious weight-related problems.

BETWEEN 25 AND 30

You may be at some risk of weight-related problems.

BETWEEN 20 AND 25

You are in the ideal zone for good health. A BMI under 20 can be linked to health risks.

John Smith is 179cm tall and he weighs 67 kg.

1. According to the chart, what is his Body Mass?
2. According to the chart, does he need to lose weight, gain weight or is his weight healthy?

(Please note: the numbers for the chart were hand-written in the hard-copy format and could not be reproduced for this electronic edition.)

Activity 5 : 次つつくテキストを読んで下さい。あたかもあなたが友だちにその内容を説明する

かのように

要約して下さい。(adapted from Maclean's Magazine 9/4/2000 pg. 32).

TERRY FOX

A few years after the tragic death of Terry Fox in 1981, a group of historians gathered to discuss the fads and fashions of the past decade – what would last and what would not. They concluded that his Marathon of Hope for cancer research had lit up a generation of people around the world, but the memories would definitely fade.

How wrong they were! Annual Terry Fox runs are held all over Canada and in 59 other countries.

After losing most of one leg to bone cancer when he was 18, Fox resolved to run across Canada to raise money to fight the disease that afflicted him. He decided to do it, in fact, just before the operation in which his right leg was amputated.

After months of preparation with his artificial leg, he set out from St. John's, Newfoundland on April 12, 1980, and ran 26 miles every day, seven days a week. However, he was forced to stop in September because the cancer had returned. In 11 months, he was dead.

The Terry Fox runs began in 1981. The first one attracted more than 300,000 participants in 760 cities and towns – in Canada and as far away as Saudi Arabia, China and the Soviet Union. Over $250 million dollars has been raised worldwide for cancer research. All because of Terry Fox.

Appendix 3 – Tape-script for TIC' Listening Tasks

Activity 1: Telephone message.

Hello, this message is for Mona Rogers. Mona, this is Kate Peters. It's Thursday the 14th, 5:30 p.m. I'm sorry for calling so late, but I can't make our meeting tomorrow as I've had to go to New York. Could you call my office and reschedule the meeting for sometime next week? Thanks very much. Bye.

Activity 2: Post office

Well, let's see… hmmm, the package weighs… 2.3 kilograms… but it's small so you could send it by airmail. It'll take approximately 10 days and it will cost $25.89.

By sea, it can take from 6 to 8 weeks, but it's less at $14.53

A middle option is SAL – that's sea and land delivery. It takes approximately 4 weeks and the charge is $19.54.

Activity 3: Map directions

O.K., take the bridge over the river and turn right onto the Parkway. Follow the Parkway to Carling Avenue. Ummm…. Be careful because the Parkway does fork – follow the road to the right. You'll come to a T-intersection… Lincoln Heights Shopping Mall is on your right.
So, turn right on Carling Avenue and then take the first right **after** Lincoln Heights. Take the next right – you'll turn onto Richmond Road. Follow Richmond and turn left on the 2nd road. You'll see a gas station on the left. My apartment building is the second one on your left.

Activity 4: The Weather report.

And now for the weather. Spring has arrived early with temperatures well above normal, and I don't think anyone will complain. We had beautiful weather today, with the temperature reaching 18C in the city, and with light winds coming in from the west. Tomorrow we can expect more of the same. Mostly sunny skies with a few cloudy periods. The morning will be cool with the temperature around 11C, but we expect to reach a high of 21C by noon. The weather should stay the same through the week, although we can expect cooler temperatures on the weekend, with some rain.

Activity 5: The Cirque du Soleil (adapted from Maclean's Magazine 9/4/2000 pg. 45)

Everybody knows the circus – they've been around for centuries, with their elephants, lions, dancing bears and funny clowns; but have you heard of the Cirque de Soleil?

The Cirque de Soleil is a new style of circus from Canada that has captured the heart of the world with its energy, imagination and raw talent.

Its origins began with the age-old tradition of street performances. In 1982, Guy Laliberte, Gilles Ste. Croix and Daniel Gauthier created a summer festival in Quebec. It featured their friends who were stilt-walkers, acrobats and other street entertainers. The festival was very successful and in 1984, they created the Cirque du Soleil. The circus first performed in Quebec and attracted huge crowds with its new style. There were no animals, only talented jugglers, contortionists who bend their bodies in the most amazing way, trapeze artists and acrobats.

After its debut success, it toured all over Canada and the United States. Cirque du Soleil was so successful that it went to Europe in 1990, and to Japan in 1992.

The circus is hugely successful – you can see it on a trip to Las Vegas or to Walt Disney World. However, the Cirque de Soleil has not lost touch with its origins, that of giving joy and happiness to people. Cirque du Soleil gives 1 percent of its box office earnings to programs that help young people stay off the streets.

REFERENCES

Aitchison, J. (1994). *Words in the Mind: an introduction to the mental lexicon*. (2nd edition). Blackwell.

Alderson, J.C. and D. Wall. (1993). "Does Washback Exist?". *Applied Linguistics* 14/2:115-129.

Arbogast, B., J. Bicknell, T. Duke, M. Locke (2000). *TOEIC Official Test-Preparation Guide.* ETS: Princeton, New Jersey.

Asahi Shimbun February 22, 2000 pg. 1. "*Eigo Damenara Shoshin Mo Dame" (*If You Are Bad at English, Your Career Will Stall).

Bachman, L.F. (1990). *Fundamental Considerations in Language Testing*. Oxford University Press.

Bachman, L.F. (2000). "Modern Language Testing at the Turn of the Century: assuring that what we count counts". *Language Testing* 17/1:1-42.

Bachman, L.F. and A. Palmer (1996). *Language Testing in Practice*. Oxford University Press.

Brazil, D. (1992). "Speaking English or Talking to People". Unpublished lecture given at Sophia University, Tokyo Japan.

Brown, H.D. (1994). *Principles of Language Learning*. Prentice Hall.

Brown, J.D. (1995). "Differences Between Norm-Referenced and Criterion-Referenced Tests". In Brown and Yamashita (eds.) pages 12-19.

Brown, J.D. and S.O. Yamashita (eds.) (1995). *Language Testing in Japan*. JALT.

Brown, J.D. and S.O. Yamashita (1995). "English Language Entrance Examinations at Japanese Universities: 1993 and 1994". In Brown and Yamashita (eds.) pages 86-100.

Brumfit, C.J. and K. Johnson (eds.) (1979). *The Communicative Approach to Language Teaching*. Oxford University Press.

Canale, M. (1983). “From Communicative Competence to Communicative Language Pedagogy”. In Richards in Schmitt (eds.) pages 2-27.

Carter, R. and M. McCarthy (eds.) (1988). *Vocabulary and Language Teaching*. Longman.

Chan, M.M. (1986). “Teaching Writing as a Process of Communication at the Tertiary Level”. *JALT Journal* 8/1:53-70.

Childs, M. (1995). “Good and Bad Uses of TOEIC by Japanese Companies”. In Brown and Yamashita (eds.) pages 66-75.

Conrad, L. (1985). “Semantic Versus Syntactic Cues in Listening Comprehension”. *Studies in Second Language Acquisition* 7:59-72.

Cook, G. (1989). *Discourse*. Oxford University Press.

Coulthard, M. (1985). *An Introduction to Discourse Analysis*. Longman.

Cunningham, C. (2003). Company Culture and Ideological Differences: An Analysis of Split Within One University in Relation to Curriculum Goals. http://www.cels.bham.ac.uk/resources/Essays.htm (1/15/2003).

de Bot, K., R. Ginsberg and C. Kramsch (eds.) (1991). *Foreign Language Research in Cross-Cultural Perspectives*. John Benjamins.

Doughty, C. and J. Williams (eds.) (1998). *Focus on Form in Classroom Second Language Acquisition*. Cambridge University Press.

Doughty, C. and J. Williams (1998). “Issues and Terminology”. In Doughty and Williams (eds.) pages 1-11.

The *Eiken* Website: http://www.eiken.or.jp/english/history/index.html

Ellis, N. (1988). “Vocabulary Acquisition: word structure, collocation, word-class, and meaning”. In Schmitt and McCarthy (eds.) pages 122-139.

Ellis, R. (1990). “Communicative Competence and the Japanese Learner”. *JALT Journal* 13/2:103-128.

Ellis, R., H. Basturkmen, S. Loewen (2001). “Learner Uptake in Communicative ESL Lessons”. *Language Learning* 51/2:281-318.

Garfinkel, H. (1970). “Remarks in Ethnomethodology”. In Gumperz and Hymes (eds.)

Gates, S. (1995). “Exploiting Washback from Standardized Tests”. In Brown and Yamashita (eds.) pages 101-106.

Gilfert, S. (1995). “A Comparison of TOEFL and TOEIC”. In Brown and Yamashita (eds.) pages 76-85.

Gilfert, S. (1996). “A Review of TOEIC”. *The Internet TESL Journal* 2.8. http://iteslj.org/Articles/Gilfert-TOEIC.html.

Gorsuch, G.G. (1997). “Test Purposes”. *The Language Teacher Online*. http://langue/hyper.chubu.ac.jp/jalt/pub/tlt/97/jan/gorsuch.html.

Gorsuch, G.G. (1998). “*Yakudoku* EFL Instruction in Two Japanese High School Classrooms: an exploratory study”. *JALT Journal* 20/1:4-32.

Gorsuch, G.G. (2000). “EFL Educational Policies and Educational Cultures: influences on teachers’ approval of communicative activities”. *TESOL Quarterly* 34/4:675-710.

Gumperz, J.J and D. Hymes (eds.) (1970). *Directions in Sociolinguistics*. Holt Rinehart and Winston.

Hadley, G. (1998). “Returning Full Circle: a survey of EFL syllabus designs for the new millennium”. *RELC Journal* 29/2:50-67.

Hamp-Lyons, L. (1998). “Ethical Test Preparation Practice: the case of the TOEFL”. *TESOL Quarterly* 32/2:329-337.

Hamp-Lyons, L. (1999). “Polemic Gone Astray: a corrective to recent criticisms of TOEFL preparation – the author responds”. *TESOL Quarterly* 33/2:270-274.

Hayashi, K. (1995). “Form-Focused Instruction and Second Language Proficiency”. *RELC Journal* 26/1:95-115.

Hemingway, M.A. (1999). *English Proficiency Tests: a comparative study*. The Chauncey Group International.

Hilke, R. and P. Wadden. (1997). “The TOEFL and its Imitators: analyzing the TOEFL and evaluating TOEFL-prep texts”. *RELC Journal* 28/1:28-53.

Hilke, R. and P. Wadden (1999). “Polemic Gone Astray: a corrective to recent criticisms of TOEFL preparation”. *TESOL Quarterly* 33/2:263-270.

Hino, N. (1989). “*Yakudoku*: Japan’s Dominant Tradition in Foreign Language Learning”. *JALT Journal* 10/1-2:45-55.

Hughes, A. (1989). Testing for Language Teachers. CUP.

Hughes, A. and C. Lascaratou (1982). “Competing Criteria for Error Gravity”. *ELT Journal* 36/3:175-182.

Hulstijin, J. and B. Laufer (2001). “Some Empirical Evidence for the Involvement Load Hypothesis in Vocabulary Acquisition”. *Language Learning* 51/3:539-558.

Hymes, D.H. (1979). “On Communicative Competence (extracts)”. In Brumfit and Johnson (eds.) pages 5-26.

John, V. (1967). “Communicative Competence of Low-income Children: assumptions and programs”. *Report of Language Development Study Group*. Ford Foundation.

Just, M.A. and P.A. Carpenter (1992). “A Capacity Theory of Comprehension: individual differences in working memory”. *Psychological Review* 99/1:122-149.

Larsen-Freeman, D. and M. Long (1991). *An Introduction to Second Language Acquisition Research*. Longman.

Laufer, B. and T.S. Paribakht (1998). “The Relationship Between Passive and Active Vocabularies: effects of language learning contexts”. *Language Learning* 48/3:365-391.

Leonard, J. (1998). “Japanese University Entrance Examinations: an interview with Dr. J.D. Brown”. *The Language Teacher Online*. http://langue/hyper.chubu.ac.jp/jalt/pub.tlt/98/mar/Leonard.html

Lewkowicz, J.A. (2000). “Authenticity in Language Testing: some outstanding questions”. *Language Testing* 17/1:43-64.

Long, M. (1991). “Focus on Form: a design feature in language teaching methodology”. in de Bot *et al* (eds.) pages 39-52.

Long, M. and P. Robinson (1995). “Focus on Form: theory, research and practice”. In Doughty and Williams (eds.) pages 15-41.

Maclean’s Magazine January 11, 1999 pg. 57. “Using the BMI Chart”.
September 4, 2000 pg. 32. “The Runner Terry Fox”.
September 4, 2000 pg. 45. “The Entertainers Cirque du Soleil”.

Mainichi Shimbun March 30, 2000 pg. 21. *“Eigo Ga Nigeru Shusei No Kagi”* (Taking Command of English – The Key to Success).

McGregor, L. (2002). “A Student Guide to Plagiarism”. *The Language Teacher* 26/1:11-15.

McCarthy, M. (1990). *Vocabulary*. Oxford University Press.

McMurray, D. (2002). “A(nother) Student Guide to Plagiarisms”. *The Language Teacher* 26/5:18.

Mombusho (1999). *The Course of Study for Foreign Languages at Junior High Schools*. Tokyo Shoseki.

Moritoshi, P. (*nd*). “Evaluation of and Practical Suggestions for Improving a Typical Japanese Junior High School English as a Foreign Language Syllabus: a case study”. Unpublished paper as part of the University of Birmingham’s ODL M.A. Program in TEFL/TESL.

Moritoshi, P. (2003). The Test of English for International Communication (TOEIC): necessity, proficiency levels, test score utilization and accuracy. http://www.cels.bham.ac.uk/resources/Essays.htm (1/15/2003).

Morrow, K. (1979). “Communicative Language Testing: Revolution or Evolution?”. In Brumfit and Johnson (eds.) pages 143-157.

Nagy, W. (1988). “On the Role of Context in Vocabulary Learning”. In Schmitt and McCarthy (eds.) pages 64-83.

Nikkei Shimbun October 10, 2000 pg 15. *“Eigo Kenshou Ni Penaruti”* (Penalties at Company English Camps).

O’Malley, J.M., A.U.Chamot, L. Kupper (1989). “Listening Comprehension Strategies in Second Language Acquisition”. *Applied Linguistics* 10/4:418-437.

Owen, C., J. Rees, S. Wisener (1997). *Testing*. Centre For English Language Studies: The University of Birmingham.

Palmer, A.S. (1978). “Measures of Achievement, Communication, Incorporation and Integration for Two Classes of Formal EFL Learners”. *Paper Read at the 5th AILA Congress*, Montreal, August. Mimeo.

Rea, P.M. (1978). “Assessing Language as Communication”. *MALS Journal*, New Series No. 3, University of Birmingham.

Richards, D. (1980). “Problems in Eliciting Unmonitored Speech in a Second Language”. *Interlanguage Studies Bulletin* 5:63-98.

Richards, J.C. and T. Rogers (2001). *Approaches and Methods in Language Teaching. Second Edition*. Cambridge University Press.

Richards, J.C. and R. Schmidt (eds.) (1983). *Language and Communication*. Longman.

Robb, T.N. and J. Ercanbrack (1999). “A Study of the Effect of Direct Test Preparation on the TOEIC Scores of Japanese University Students”. *TESL-EJ* 3/4:1-22. http://www.kyoto-su.ac.jp/information

Rutherford, W. (1987). *Second Language Grammar: learning and teaching*. Longman.

Sankei Shimbun June 16, 1999 pg 25. *“Eigo Wa Shusei No Nishu Kamoku”* (English is a Core Employment Requirement).

Schmitt, N. (1998). “Tracking Incremental Acquisition of Second Language Vocabulary: a longitudinal study”. *Language Learning* 48/2:281-317.

Schmitt, N. (1999). “The Relationship Between TOEFL Vocabulary Items and Meaning: Association, Collocation and Word-class Knowledge”. *Language Testing* 16/2:189-216.

Schmitt, N. and M. McCarthy (eds.) (1988). *Vocabulary: description, acquisition and pedagogy*. Cambridge University Press.

Schneider, D. (2001). “Proficiency Testing and Gain Scores: a research overview and use of the TOEFL at one Japanese college”. *Journal of Aomori Public College* 9/7:26-45.

Sinclair, J. and A. Renouf (1988). “A Lexical Syllabus for Language Learning”. In Carter and McCarthy (eds.) pages 140-160.

Smith, J. (2000). “Teaching the Test Takers”. *The Language Teacher Online*. http://langue.hyper.chubu.ac.jp/jalt/pub/tlt/00/nov/smith.html

Smith, M.L. (1991). ‘Put to the Test: the effects of external testing on teachers’. *Educational Researcher* 20/5:8-11.

Spolsky, B. (1985). “What Does it Mean to Know How to Use a Language? An essay on the theoretical basis of language testing.” *Language Testing* 2/2:180-191.

Spolsky, B. (1989). “Communicative Competence, Language Proficiency and Beyond”. *Applied Linguistics* 10/2:138-156.

Swain, M. (1998). “Focus on Form Through Conscious Reflection”. In Doughty and Williams (eds.) pages 64-81.

Swain, M. (2001). “Examining Dialogue: another approach to content specification and to validating inferences drawn from test scores”. *Language Testing* 17/1:1-42.

Tarone, E. (1988). *Variations in Interlanguage*. Edward Arnold.

Terrell, T.D. (1977). “A Natural Approach to Second Language Acquisition and Learning”. *The Modern Language Journal* 61/7:325-37.

Tenth TOEIC Client Survey 1999. The Chauncey Group International.

‘The Reporter’ TOEIC Newsletter #1-6; #8-11. Educational Testing Services.

The TOEIC Steering Committee (a). (*nd*). “Proceedings From The 35th TOEIC Seminar, December 10, 1991 in Osaka, Japan”.

The TOEIC Steering Committee (b). (*nd*). “Proceedings From the 58th TOEIC Seminar, March 11th, 1997 in Tokyo Japan”.

The TOEIC Technical Manual (*nd*, electronic edition). The Chauncey Group International.

TOEIC Examinee Handbook (1996). Educational Testing Services.

TOEIC Can-Do Guide (1998). The Chauncey Group International.

TOEIC UserGuide (1999). The Chauncey Group International.

TOEIC Report on Test-Takers Worldwide 1997-1998 (2000). The Chauncey Group International.

TOEIC Test of English for International Communication: HISTORY AND STATUS. (*nd*) The TOEIC Steering Committee.

Weir, C. (1990). *Communicative Language Testing*. Prentice Hall.

Wesdorp, H. (1982). *Backwash Effects of Language Testing in Primary and Secondary Education*. Stichting Centrum voor Onderwijsonderzoek van de Universiteit van Amsterdam.

White, L., N. Spada, P.M. Lightbrown, L. Ranta (1991). “Input Enhancement and L2 Question Formation”. *Applied Linguistics* 12/4:416-432.

Widdowson, H.G. (1975). “Directions in the Teaching of Discourse”. In Brumfit and Johnson (eds.) pages 49-60.

Widdowson, H.G. (1979). “The Teaching of English as Communication”. In Brumfit and Johnson (eds.) pages 122-142.

Widdowson, H.G. (1989). “Knowledge of Language and Ability For Use”. *Applied Linguistics* 10/2:128-137.

Williams, J. (1999). “Learner-Generated Attention to Form”. *Language Learning* 49/4:583-625.

Willis, D. (1990). *The Lexical Syllabus*. Collins Cobuild.

Woodford, P.E. (1978). “The Test of English for International Communication (TOEIC)”. presented December 19, 1978 at KEIDANRAN, Tokyo, Japan.

Woodford, P.E. (1982). *An Introduction to the TOEIC: The Initial Validity Study*. ETS, Princeton, New Jersey.

Wu, Y. (1998). “What do Tests of Listening Comprehension Test? – a retrospection study of EFL test-takers performing a multiple-choice task”. *Language Testing* 15/1:21-44.

Queries may be addressed to the author at the following e-mail address iam1smartcookie@hotmail.com