A Comparability Study on SNUCREPT and TOEIC

Inn-Chull Choi

The present study attempts to investigate the extent to which TOEIC is valid and is comparable to SNUCREPT. The data of the two tests results were obtained from approximately 300 subjects who are considered to represent the target test-takers of TOEIC and SNUCREPT. The quantitative and qualitative analyses of the test results indicate some of the validity problems with TOEIC in the content and the test method facets. The content and the factor analyses also reveal that the two tests are not comparable to each other due to the significant discrepancy between the Parts Ⅰ, Ⅴ, Ⅵ, Ⅶ of TOEIC and the counterparts of SNUCREPT in terms of the content topic and the test methods.

I. Introduction

1.1. Purpose

TOEIC (Test of English for International Communication) developed by ETS under the sponsorship of Japanese corporations, is said to be designed to measure communication skills. The test appears to have been accepted as the desirable model for measuring communicative competence. However, no systematic research report has been published as to the validity and the reliability of TOEIC. Given the great demand for a valid tool for measuring genuine communicative competence, it may well be worth the effort to compare the test content and the test method facets of TOEIC and SNUCREPT (Seoul National Univ. Criterion-Referenced English Proficiency Test). As a follow-up study on the construct validation of SNUCREPT, the current research is intended to investigate the extent to which TOEIC is valid and to explore the possibility and plausibility of equating TOEIC with SNUCREPT. Such an effort will shed light on how to develop a more valid tool to achieve the goal of measuring genuine communicative competence.

1.2. Research Methodologies

Vadid data were obtained from approximately 300 students from four different universities (including Seoul National Univ., Korea Univ., Chunnam National Univ., and Sungshin Women's Univ.). The consideration of test security limited the number of participating schools. Except for Seoul National Univ. and Chunnam National Univ., the participating subjects were students majoring in English teaching and/or English literature. Though limited in number, these universities are believed to represent a wide range of student ability in terms of admission standards. The test-taking sample may well be representative of the target test-takers, who are required to demonstrate their ability to communicate in English for their careers.

To investigate the extent to which TOEIC is valid in terms of TMF (Test Method Facet: Bachman 1990) and content validity, the present study employs several quantitative statistical analyses such as a dimensionality check based on Stout-based Factor Analyses, a correlational study, and qualitative distractor analyses. Based on the dimensionality check, the possibility of test equating between the two tests is explored through regression or IRT true-score equating. PC-BILOG and MicroCat are used to estimate the test/item and ability indices.

2. Descriptive Statistics

Note: Hereafter, CR4 refers to the SNU Criterion-Referenced 4th Pilot Test. 4L, 4G, 4V, and 4R denote CR4 Listening Comprehension test, Grammar test, Vocabulary test, and Reading Comprehension test, respectively.

As for TOEIC, T1, T2, T3, T4, T5, T6, and T7 represent TOEIC Part I (One Picture), Part II (Paraphrasing), Part III (Short Conversation), Part IV (Short Talk), Part V (Gap-Filling), Part VI (Error Recognition), Part VII (Reading Comprehension), respectively.

2.1. CR4 Item Statistics

The skewness indices of Table 1 indicate that the listening comprehension, grammar and vocabulary tests prove somewhat difficult while the reading comprehension test turns out to be a bit easy. The fact that the means are almost identical shows that the shapes of distribution are not far

Table 1. CR4 Item Statistics

	4L	4G	4V	4R
N of Items	60	50	50	40
N of Examinees	242	244	245	241
Mean	28.239	23.840	22.352	21.660
Variance	56.633	40.426	40.559	26.990
Std. Dev.	7.525	6.358	6.369	5.195
Skew	0.362	0.198	0.390	0.014
Kurtosis	0.178	−0.329	−0.313	0.301

Table 2. TOEIC Item Statistics

	T1	T2	T3	T4	T5	T6	T7
N of Items	20	30	30	20	40	20	40
N of Examinees	282	282	282	282	282	282	282
Mean	9.011	12.677	13.583	7.110	23.493	9.550	20.163
Variance	8.230	20.729	25.134	7.233	32.165	14.850	36.654
Std. Dev.	2.869	4.553	5.013	2.689	5.671	3.854	6.054
Skew	0.104	0.607	0.024	0.296	−0.493	0.105	−0.141
Kvurtosis	−0.143	0.847	0.075	0.365	−0.093	−0.663	−0.139

from symmetrical distribution. As for the kurtosis indices, the vocabulary and grammar tests have a more flat distribution (which is ideal for a competitive large-scale test) than the reading and listening comprehension tests. The above indices in general suggest that the performance of the sample test-takers does not seriously violate the assumption of normal distribution.

2.2. TOEIC Item Statistics

The skewness indices of Table 2 ranging from −0.493 to 0.607 reveal that Part II is a little difficult (positively skewed) whereas Part V is a bit easy (negatively skewed). The kurtosis indices ranging from −0.663 to 0. 847 indicate that Part II has a slightly peak-shaped distribution (which is not desirable for a large-scale distribution) and that Part VI has a somewhat flat distribution. The distributions in general have the shape of a fairly normal distribution.

2.3. CR4 Test Statistics

Table 3 shows that the overall difficulty is around .5, which reveals that

Table 3. CR4 Test Statistics

	4L	4G	4V	4R
Alpha	0.787	0.758	0.784	0.753
SEM	3.472	3.128	2.960	2.784
Mean P	0.471	0.477	0.447	0.541
Mean Biserial	0.354	0.366	0.391	0.379

Table 4. TOEIC Test Statistics

	T1	T2	T3	T4	T5	T6	T7
Alpha	0.520	0.703	0.778	0.439	0.778	0.737	0.788
SEM	1.988	2.481	2.362	2.014	2.672	1.975	2.788
Mean P	0.451	0.423	0.453	0.355	0.587	0.477	0.504
Mean Biserial	0.419	0.415	0.475	0.378	0.449	0.529	0.423

the difficulty level set by the test developers corresponds to the overall ability level of the target test-takers. The overall discriminability is over .35 which is considered adequate for a large-scale test. The Cronbach Alphas, the reliability indices for all the tests are higher than .753, which is considered to meet the criterion of a large-scale standardized test.

2.4. TOEIC Test Statistics

Table 4 shows tat the overall difficulty level ranges from .355 to .504, which is a little lower than the mean of normal distribution. This suggests that the difficulty level of the test designed for the world-wide market may not be appropriate for culture-specific test-takers. Given the fact that the difficulty level of a test which is not appropriately set for the test-takers leads to lower reliability, the inherent problem of the difficulty level should be considered seriously in using an international test.

The overall discriminability level ranges from .378 to .529, which is considered adequate. Considering that Parts 1 & 2 have a lower reliability than Part VI which has the same number of test items (20), the content validity should be investigated to account for their relatively low reliability indices. The overall reliability indices of TOEIC prove to be as high as CR4 in the range between .7 and .8.

3. Test Method Facets of TOEIC

TOEIC is composed of two sections, the listening and the written sections.

The listening section consists of four parts including: 1) Picture Description (20 question items); 2) Question and Response (30 items); 3) Short Conversation (30 items); and 4) Short Talks (20 items). The written section is made up of three parts including 5) Grammar and Vocabulary in Sentence Completion Task (40 items); 6) Error Recognition (20 items); and 7) Reading Comprehension (40 items). Given such structured test methods of TOEIC, the present paper attempts to explore the problematic areas of TMFs (Test Method Facets) based on quantitative analyses as well as qualitative analyses based on theories of language use and the recent findings of the SNUCREPT research (Choi, 1994).

3.1. Part I

3.1.1. Problems with Reference, Culture, and Artificiality

Item analyses suggest that the most serious inherent problem is the vague description of a given picture whose interpretation can vary from test-taker to test-taker. This inherent problem of Part I is attributed to several factors which are explained as follows. The first factor is ambiguity of reference, that is, the unclear objects in a given picture are referenced by verbal description. This is illustrated in question # 18, in which it is difficult to reference the boy as in the choice C, "The boy standing has dropped his pants." It can either refer to the boy in the back or the one in the front. Cultural difference is the second factor. The difference between the test-taker's culture and the culture reflected by the picture can cause problems of interpretation, as is shown in questions # 7 and 15. The third factor is the inherent disparity between the motionless picture and the verbal descriptions of motion. The stationary picture fails to manifest the verbal description of motion. As is illustrated by question # 15, the picture fails to show an open flame as in the choice b., "The chef is cooking over an open flame." Additionally, the picture quality could be improved to clearly show what the picture is intended to portray.

<Item # 7>

Alt.	P/E	r_{bis}	Key
1	0.408	0.020	*
2	0.337	0.378	?
3	0.011	−0.418	
4	0.241	−0.396	
Other	0.004	−1.000	

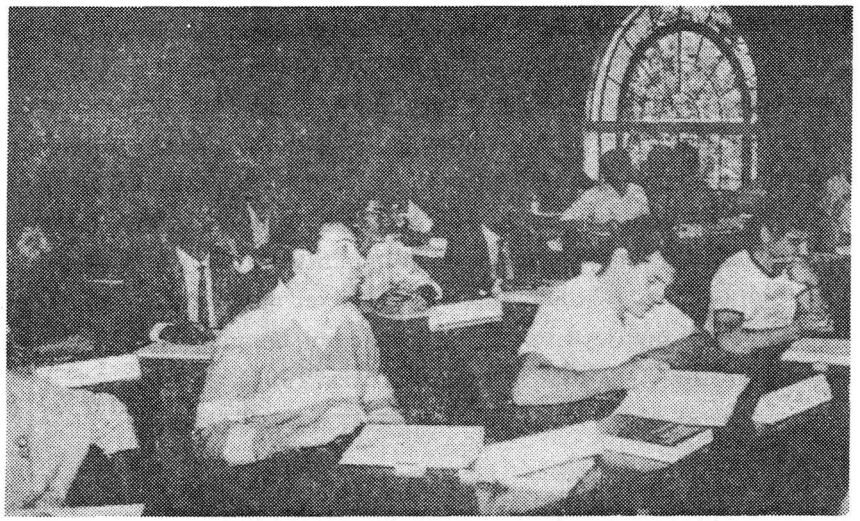

Fig. 1. Item # 7

- Alt. represents Alternative.
- P/E represents Proportion Endorsing.
- r_{bis} represents biserial correlation.
- Under Key, * denotes an answer key and '?' denotes a distractor that has a higher discriminability index than the key.

<Item # 15>

Alt.	P/E	r_{bis}	Key
1	0.067	−0.256	
2	0.465	−0.062	*
3	0.376	0.126	?
4	0.089	0.128	
Other	0.004	−1.000	

<Item # 18>

Alt.	P/E	r_{bis}	Key
1	0.184	0.037	
2	0.255	−0.344	
3	0.209	0.235	?
4	0.348	0.119	*
Other	0.004	−1.000	

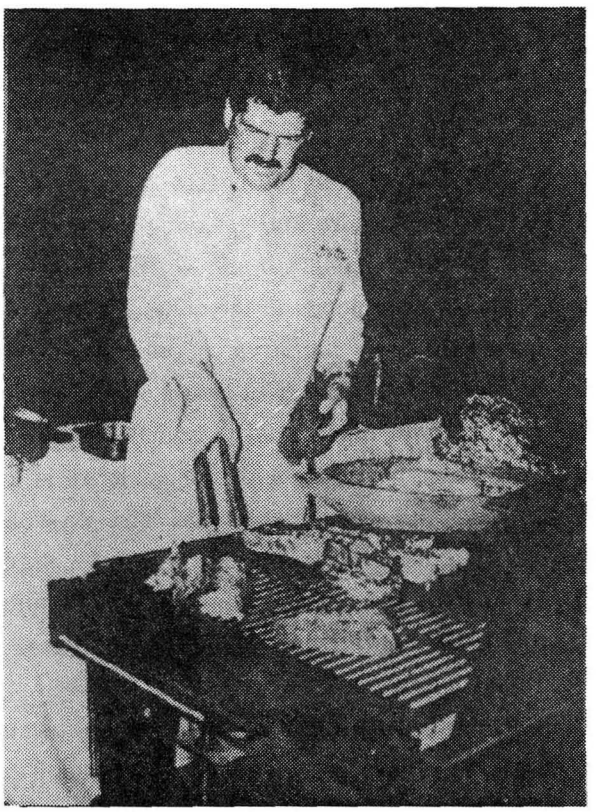

Fig. 2. Item # 15

Fig. 3. Item # 18

3.1.2. Effect of Field-In/Dependence on Picture Description

In Part I, test-takers are supposed to listen to three choices while looking at a given picture and choose the choice which best describes the picture. It appears to be a valid method to measure aural comprehension within a visual context. In fact, however, many test-takers have difficulty identifying the details in a given picture and grasping the bits and pieces of information conveyed by the visual stimuli while almost simultaneously choosing the choice which corresponds to the content of the picture.

Without consciousness of their cognitive process, test-takers are expected to resort to their test-taking strategies based on an important factor of language acquisition, i.e., field-independence and field-dependence. The inclination toward field-dependence or field-independence is known to vary from individual to individual at different ages. The external influence of such a non-linguistic cognitive style factor on aural comprehension test results should be minimized. The relatively low discriminability is indicative of the fact that such a non-linguistic factor has a significant effect on test performance.

Visual clues can be used as an added context in the test booklet, rather than as an essential part of the question itself. This type of indirect presentation of visual clue can be complemented by utilizing the VCR system. To supply context-rich visual clues can make the test method more valid for measuring aural comprehension skills.

3.2. Part II

3.2.1. Context-Reducedness: Function Words vs. Content Words

Part II requires test-takers to choose the best paraphrased version of a listening stem. This task is lacking in context in communication in that the stem which consists basically of one single short sentence, which is likely to provide an inadequate context. It is important to try to provide a rich context within a single sentence stem. The mere advantage of such a relatively simple task has more discriminating power between upper and lower ability groups.

With such context-reduced tasks as paraphrasing, it is also important not to include items which measure the ability to recognize the weakened sounds of function words. As is demonstrated in question # 44 and # 46,

low discriminability is attributed to assessing the ability to recognize the weakened consonant/vowel of a function word [for example "her" as in "How did the doctor describe her condition?", "they" as in "Didn't he say they're going to be here at 4:30?" in a context-reduced sentence]. In a context-reduced sentence, it is not wise to force the test-takers to focus their auditory perception on function words which have weakened vowels or consonants. This violates the idea of a real-life communication setting, in that context allows listeners in general to base their aural comprehension processes on content words and guess function words within a context. In Part II, it would be wise to assess the auditory recognition skills by focusing more on content words than on function words.

<Item # 44>

Q: How did the doctor describe her condition?

a. He said I was in perfect health.

b. He turned it down because of all the conditions.

c. She said it was serious, but that she was optimistic.

Alt.	P/E	r_{bis}	Key
1	0.404	0.351	?
2	0.230	−0.400	
3	0.348	0.075	*
Other	0.011	−1.000	

<Item # 46>

Q: Didn't he say they were going to be here at 4:30?

a. No, she didn't.

b. That's right. He didn't.

c. I wonder where they are.

Alt.	P/E	r_{bis}	Key
1	0.507	0.195	?
2	0.259	−0.188	
3	0.216	0.077	*
Other	0.011	−1.000	

3.2.2. Limited Number of Choices

The number of choices in Part II is limited only to three, which makes the probability of getting an item correct by chance as high as 33.3%. Little rationale is given for providing only three choices, i.e., only two distractors and one key.

It is self-evident that providing four choices produces more reliable results by lowering the probability of getting an item correct by chance. The memory load problem of listening to four choices can be solved by presenting choices consisting of short meaningful phrases or sentences (rather than by giving three choices). With four shorter choices, the possibility of guessing can be lowered, and discriminability can be enhanced.

3.3. Parts III & IV

3.3.1. Test-Taking Strategies

The problematic area of the test method facets of Part III and IV lies in the fact that test-takers listen to the short conversation and short talks, look at the questions and choices written in the test booklets, and then answer them. This method leads to less reliable results in that the test performance depends on which test-taking strategy a test-taker chooses to employ. That is, a test-taker may wish to look at the question and the choices first and then listen to the aural comprehension passage, or listen to the passage first and then look at the question and the choices later, or engage in the two tasks simultaneously. Whichever strategy he or she chooses to employ, the test results will eventually vary with lowered reliability.

3.3.2. Interpretation of Test Results

This method which combines spoken-mode and written-mode English causes a dimensionality problem. The listening comprehension test with choices in written form requires the test-takers to activate their aural comprehension skills adulterated with the ability to decode written English. Such a task inevitably forces the test-takers to resort to the di-dimensional abilities to decode spoken and written form of language. The two separate abilities represented by one score makes it virtually impossible to interpret the scores appropriately.

3.3.3. Artificiality of Question Content

There are questions which require test-takers to infer the place where a given conversation is taking place, i.e. "Where are they?" (as in item # 53, 61, 74, 80 presented in Appendix A) or "Where is this talk being given?" (as in item # 94 presented in Appendix A). Such question content, however, is most problematic in that its artificiality most seriously violates the fundamental assumption of real life communication. That is, we can hardly think of a single situation in a natural setting in which we are required to infer the kind of place we are located in. (Unless we are forced to be blindfolded and to infer where we are located in a kidnapping incident!)

This issue can be easily solved only if we can utilize video facilities for listening tests in which test-takers are provided with all the visual clues describing the circumstances in motion pictures. Under the circumstances, however, presenting a listening test through an audio and video channel may entail too many technical problems in an ordinary testing environment. The possible utilization of video facilities for listening tests merits further research. This thorny problem can be indirectly solved by allowing the test-takers to be exposed to the aural passage two times, as is explained in Choi (1994).

3.3.4. Exposure Time to the Aural Passage

TOEFL and TOEIC employ a test method for listening tests in which the listening passages and questions are presented once in a rather slow speech, i.e. approximately 145 WPM (Words Per Minute). This method fails to reflect the normal, natural speed of speech (160-190 WPM) used in real-life communication. The speed factor functions as an extremely important factor in producing sandhi phenomena, which pose serious problems to non-native listeners who are accustomed to listening to TESLese. The slow speed of speech plays a significantly negative role in the measurement of aural comprehension skills as well as for listening education (Choi, 1992).

As was shown in the recent research (Choi, 1994) on SNUCREPT, it is desirable to allow test-takers to listen to a listening passage and the corresponding question twice and to the articulately presented choices once. Two time exposure has long been implemented in the listening test of the Cambridge Proficiency Test Batteries of the University of Cambridge Local Examination Syndicate (UCLES), whose content validity is found to be supe-

rior to that of TOEFL (Bachman et al., 1995).

Especially in a Question and Answer format, this test method proves valid in terms of the reflection of listening processes. The test-takers listen to the passage first in a macro-listening manner, and then to the question. Only then do they understand what they are required to listen for. This procedure allows them to listen for specific information when they listen to the passage again in a manner of micro-listening. Thus, this method induces a natural way of activating both macro-listening and micro-listening skills.

3.3.5. Cultural Bias

The problematic items with low discriminability are the ones which require the test-takers to have specific background knowledge of a specific culture. Granting that the language test cannot avoid containing culture-based language content, the results of a language test should not be interpreted as the degree of cultural awareness. A foreign language test should focus on the foreign language ability through culture-embedded language-proper content, not merely on the understanding of a given foreign culture. Representing a test of English as international communication, the test is expected to cover international and culturally neutral topics. As in items # 94 and #95, for example, the test items dealing with culturally biased topics lead to lower discriminability. Furthermore, as is forementioned, the content of question #94 ("Where is this talk being given?") is invalid in terms of real-life communication setting.

<Item # 94>

Questions 94 and 95 refer to the following talk.

I'd like to welcome everyone to Pennsylvania Foods. Before we begin our tour, we'd like to point out that. Unfortunately picture-taking is not allowed inside the plant. You can check your cameras at the security counter over there. Also I'd like to remind parents to hold young children by the hand. This plant is not a playground. And much of the robotic machinery you see can be dangerous for unsupervised children. Now, I hope I haven't frightened you too much. Let's step this way and begin the tour.

Now read question 94 in your testbook and answer it

Q: Where is this talk being given?

a. At a food processing plant
b. At an amusement park
c. At a camera manufacturer
d. At a robot manufacturing plant

Alt.	P/E	r_{bis}	Key
1	0.220	0.277	?
2	0.287	−0.123	
3	0.213	−0.141	
4	0.266	0.078	*
Other	0.014	−0.692	

<Item # 95>
Now read question 95 in your testbook and answer it

Q: This announcement warns AGAINST
a. picture taking in the security area
b. allowing children to move about freely
c. touching the robots
d. allowing children outside the security area

Alt.	P/E	r_{bis}	Key
1	0.273	0.181	?
2	0.298	0.047	
3	0.135	−0.252	*
4	0.270	0.031	
Other	0.025	−0.549	

3.4. Part V

3.4.1. Distribution of Grammatical Points

Item analyses suggest that Part V of TOEIC is problematic in that little consideration appears to have been taken as to the distribution of grammatical categories. More consideration should have been given to balance the distribution of grammatical points, which are as follows.

Determiner (101); Noun (128); Pronoun (107, 111, 124); Rel. Pronoun /Adjective (118, 120); Verb (104, 105, 115); Auxiliary Verb (117); Gerund (108, 125); Infinitive (140); Past Participle (134, 139); Adjective

(102, 113, 122, 129, 136); Adverb (123, 138); Preposition (109, 121); Conjunction (112, 114, 116, 132); Subjunctive Mood (119, 135); Voice (110); Basic Structure (126); Tense (137); Inversion (133); Vocabulary (103, 106, 127, 130, 131). (Due to the limited space, the questions will not be provided in this paper.)

Items on the discrete-point grammatical points such as adjectives should have been given in the category of more integrative grammatical points such as a basic sentential structure, which constitutes an essential component of communicative competence.

3.4.2. Combination of Grammar and Vocabulary

There are pros and cons to presenting two linguistic components — grammar and vocabulary — in the same format on the same test. The advantage may be that the test-takers are less likely to activate their explicit linguistic knowledge called learning. This gap-filling task part is slightly more desirable test method than the TOEFL Section 2 Structure test format, in which only four parts (one word or a combination of words) are underlined, but the other parts are excluded as choices.

As was explained in Choi (1994), the effect of this, however, can be maximized only when the test is conducted in the format of a speeded test. The disadvantage is that the combination of two linguistic components makes it impossible to interpret the test scores to evaluate the ability of the test-takers pertaining to the respective content domain.

3.5. Part VI

3.5.1. Artificiality of the Discrete-Point Test

The test method of error recognition among the underlined parts of a given sentence, a method which is of the discrete-point test type, fails to reflect real-life language use. Such a discrete-point test is not considered valid for proficiency tests. It also prompts the test-takers to rely heavily on explicit knowledge of grammatical points, rather than to approach the study of grammar in a more global manner (Oller, 1979; Savignon, 1986). Such a test method encourages language learners to build up the explicit grammatical knowledge — the core of the learning (Krashen, 1985) which is known to impede the acquisition of genuine communicative competence.

This test method which forces students to build up mechanical skills is

known to have relatively high discriminability in that there tends to be a great discrepancy in performance between those who have prepared for the test method and those who have not. The inherent problem with the discrete-point oriented tasks is well documented in a TOEFL Research Report (Henning and Cascallar, 1992) which reveals that Written Expression of TOEFL fails to have a high correlation with measures of communicative competence.

3.5.2. Context-Embedded Test Method

This method can be improved by dividing a sentence into four parts with three slashes. Such a method is valid in that test-takers are expected to look at grammar in a more macroscopic fashion. The method can be improved to provide more context which allows for the use of implicit knowledge of grammar. Thus, the revised method provides a more context-embedded format, i.e., a slightly longer discourse (for written English) or a two-exchange dialogue (for spoken English). In a discourse type format, the clause should be considered a basic meaningful chunk for parts separated with slashes.

3.6. Part Ⅶ

3.6.1. Biased Reading Topics

Recent research has corroborated the fact that the reading passage factor plays a most significant role in the test performance of test-takers in that their overall comprehension depends heavily on their background knowledge of the topic of the reading materials (Choi, 1992). Granting that the test is mainly designed to measure communication skills, the topics of the reading passages in the reading comprehension test are too biased toward so-called practical or business-oriented fields/concepts.

The notion of ESP appears to have been misinterpreted in the name of developing a practical or business-oriented English test. It is common sense that all kinds of business is conducted not only formally in the office but also informally in private. More often than not, informal English skills outweigh professional and formal English at a negotiation table. Competent businessmen should be equipped with the ability to handle a variety of topics and linguistic dimensions of communicative competence in terms of context-embeddedness and cognitive demand. In order to conduct international

business successfully by maintaining good human relation among international colleagues, businessmen are expected to have a good command of English covering a wide variety of topics such as news, liberal arts, non-esoteric academic readings as well as business-related reading materials. A desirable test should be geared towards measuring global comprehension skills which require a good grasp of background knowledge on a wide array of topics.

3.6.2. Specialized Terms: Jargon

Some of the biased reading passages (as in Questions 164, 176, 181., see below) contain too many specialized terms. Such jargon of too specific and/or technical a topic requires test-takers to be equipped with background knowledge of the given topic. Despite the fact that background knowledge constitutes the most fundamental basis of the complex reading processes, it is important to maximize the validity and reliability of the test by minimizing the unwarranted influence of background knowledge or technical terms on the reading comprehension test. It is clear that the test results should be interpreted not as a function of the test-taker's background knowledge but as the performance mainly of their English reading comprehension skills.

<Item # 164>

Systems Requiring Modules

If your system is described in the table below, you must apply the additional protection specified

Cable type-Application/Use	Connectors	Module
System with a modem or FAX		
Telephone (single line)	RJ11 modular jacks	M-Telr
System connected to data and/or printer networks		
Twisted pair-10BaseT/StarLAN	RJ45 modular hacks	M-TUPC
Twisted pair-Token Ring/ARCNET		RJ11/45
modular jacks	M-TR6D	
This coax-BusLANs	BNC-T adapter	M-TVD
Coax-StarLANs & IBM 3277	BNC coax connectors	M-359D
RS232 cable	DB9 connectors	M-RS2X
Systems connected to TV lines		
Cable TV	F-coax connectors	M-CATV
Flat TV Antenna	Flat TV antenna wire	M-FLTV

Which type of module can be used together with RJ11 modular jacks?

a. Only the M-TELR

b. The M-TUPC and M-TR6D

c. The M-TELR and M-TR6D

d. The M-TELR, M-TUPC, and M-TR6D

Alt.	P/E	r_{bis}	Key
1	0.553	0.188	?
2	0.077	0.158	
3	0.289	−0.155	*
4	0.060	−0.015	
Other	0.021	−0.841	

<Item # 176>

CITY	SKY	LO	HI
Chicago	sn	−3	1
Denver	sf	−10	0
Honolulu	s	21	25
Los Angeles	s	11	23
Miami	pc	21	22
New Orleans	t	23	25
New York	r	−6	−4
San Francisco	pc	7	17

Legend: s-sunny, pc-partly cloudy, c-cloudy, sh-showers, t-thunderstorms, r-rain, sf-snow flurries, sn-snow, i-ice

In which city is the weather likely to be mildest?

a. Los Angeles

b. Miami

c. New Orleans

d. San Francisco

Alt.	P/E	r_{bis}	Key
1	0.493	0.148	?
2	0.285	0.027	*
3	0.102	0.063	
4	0.074	−0.127	
Other	0.046	−0.640	

<Item # 181>

Recipe of the Month Bilbery Jam

The berries can be used to make a delicious jam, using one pound of preserving sugar to each pound of fruit. It is better to mix bilberries with other sour fruits (apples, cranberries, rhubarb), to improve the setting quality of the jam.
Boil the fruit with the sugar and just enough water to cover for about 20 minutes, or until the jam sets when tested on a cool saucer.
Put into clean pots and seal down.

This recipe was probably published in
a. a magazine.
b. a newspaper.
c. a cookbook.
d. a regional guidebook.

Alt.	P/E	r_{bis}	Key
1	0.162	0.130	*
2	0.025	−0.008	
3	0.683	0.113	?
4	0.035	−0.495	
Other	0.095	−0.197	

3.6.3. Simplistic Reading Skills

More often than not, it is easy to understand the main idea of the majority of practical reading passages. Therefore, too many questions focus predominantly on such comparatively simplistic skills as scanning and identifying detail information (as in 162, 163, 164, 167, 169, 170, 172, 174, 176, 177, 179, 180, 182, 183, 185, 187, 189, 191, 192, 194, 195, 196, 197, 198, 200. Due to the limited space, the questions will not be provided in this paper).

On the other hand, the reading comprehension test fails to measure global reading skills including a grasp of main ideas, coherence, and inference. Therefore, the test content is judged to be biased in measuring reading com-

prehension subskills.

3.6.4. Common Sense Questions

There are some questions dealing with practical reading passages (as in 193, 195 presented in Appendix A) in which common sense plays a crucial role in solving the reading comprehension test items. It is a serious problem that test-takers can get the item correct based only on their common sense regardless of their English reading comprehension skills. Every effort should be made not to make an English test a common sense quiz.

3.7. Local In/Dependence

All the parts except Parts Ⅳ and Ⅶ maintain a rule that one problem has one question item based on one stem or passage, which corresponds to the principle of local independence. Parts Ⅳ and Ⅶ, however, have a test method in which each of the listening and reading passages is followed by two to three question items. The two or three questions are interrelated with respect to the given passage. This fact violates the strong assumption of IRT, local independence, which means that the probability of getting an item correct should be independent of the probability of getting any other item correct. Therefore IRT cannot be applied to Part Ⅶ. For the appropriate application of IRT, the one-passage-one-item principle should be maintained (Choi, 1994).

Moreover, as was pointed out in the previous research (Choi, 1994), the three questions attached to the lengthy listening passage make it difficult even for native speakers to concentrate and memorize all the details presented in the passage. A question whose test method burdens even native speakers with too much memory load cannot be considered a desirable test item designed for non-native speakers.

3.8. Score Weight

Test-takers are given 75 minutes to solve 100 questions for the written part, which includes Parts Ⅴ (sentence completion task of 40 items), Ⅵ (error recognition task of 20 items), and Ⅶ (reading comprehension task of 40 items). There is no restriction as to the time limitation for each part. Thus, each item is expected to carry the same weight regardless of the rela-

tive difficulty of the different content domains and tasks. This clearly violates the self-evident logic that the reading comprehension skills involve more complicated cognitive processes than the linguistic components such as vocabulary and grammar. To enhance validity and reliability, the more complicated skills should be given more testing time and subsequently more score weight than the less complicated linguistic component.

Furthermore, failure to restrict the testing time on the vocabulary and grammar tests promotes the monitor use exploiting the linguistically explicit knowledge called 'learning' as opposed to 'acquisition' (according to Krashen's dichotomy). The speeded test has been shown to maximize the discriminability in terms of communication skills and thus enhance the validity and reliability (Choi, 1994). The validity of the grammar test results of TOEIC (which is supposed to measure genuine grammatical competence that can be used for communicative purpose) can be called into question in further research.

4. Dimensionality Check

Dimensionality was checked to investigate the extent to which the TOEIC test is unidimensional. The Stout procedure of factor analyses shows that IRT cannot be applied to TOEIC, which violates the strong assumption of local independence. Almost every problem has two or more question items. Therefore, classical testing theory is employed, where appropriate, for the comparability study.

4.1. Dimensionality

Stout's Dimtest procedures based on factor analyses were employed to check dimensionality (which is the fundamental assumption of IRT) of test batteries of CR4 and TOEIC. The following are the summarized results of the dimensionality check procedures (Stout et al., 1991).

The CR4 test batteries prove to be essentially unidimensional based on Stout's factor analytic dimensionality check. 4G and 4V prove to be less unidimensional in that the speeded test type appears to play a significant factor. On the other hand, violation of the unidimensionality assumption is statistically significant at alpha of .05 in some parts of TOEIC, i.e., T3, T4,

Table 5. Stout's Dimtest Statistics

	T	P-value	Dimensionality
4L	.891445	.186345	Essentially Unidimensional
4G	1.183108	.056714	Essentially Unidimensional
4V	1.592964	.055584	Essentially Unidimensional
4R	−.074684	.529767	Essentially Unidimensional
T1	−1.234890	.891564	Essentially Unidimensional
T2	.254146	.399691	Essentially Unidimensional
T3	1.956174	.025222*	Multidimensional
T4	1.522678	.030599*	Multidimensional
T5	1.419890	.017820*	Multidimensional
T6	.793948	.213613	Essentially Unidimensional
T7	−.174317	.569192	Essentially Unidimensional

*: multidimensional at alpha −.05

and T5. Such multidimensionality of T5 can be attributed to the combination of grammar and vocabulary test items. T3 and T4 fail to meet the essential unidimensionality assumption in that both have di-modal linguistic input, i.e., aural stems (dialogues and listening passages) and visual clues (questions and choices). (For reference, see the factor analyses results in the Appendix B.)

The forementioned findings strongly suggest that the data of the CR4 test batteries can be analyzed with IRT, while those of the TOEIC test batteries cannot be analyzed within the IRT framework. The discrepancy of dimensionality between the two tests invalidates the use of IRT-based true score equating.

5. Comparability of CR4 and TOEIC

5.1. Correlations between CR4 & TOEIC & Communicative Competence Measure

The number of subjects constituting valid cases for the study was limited to 42 due to the fact that the valid data could be collected only from the test-takers who took both TOEIC and CR4—and a series of tests of overall communicative competence (CC). For maximum reliability, the measures of

Table 6. Correlations between CR4 and TOEIC and CC

Cor:	L4	G4	V4	R4	T1	T2	T3	T4	T5	T6	T7	CC
L4	1.0000	–	–	–	–	–	–	–	–	–	–	–
G4	.7623^	1.0000	–	–	–	–	–	–	–	–	–	–
V4	.6563^	.6907^	1.0000	–	–	–	–	–	–	–	–	–
R4	.6070^	.4986^	.3902	1.0000	–	–	–	–	–	–	–	–
T1	.2744	.3450	.3329	.1971	1.0000	–	–	–	–	–	–	–
T2	.5307^	.6172^	.5989^	.4019*	.3931	1.0000	–	–	–	–	–	–
T3	.6583^	.6026^	.5165^	.3827	.2486	.6366^	1.0000	–	–	–	–	–
T4	.3371	.1641	.4106*	.1349	.1146	.2612	.2678	1.0000	–	–	–	–
T5	.4865*	.5326^	.3545	.5605^	.0707	.4416*	.4549*	.1996	1.0000	–	–	–
T6	.4555*	.5958^	.3643	.4385*	.1714	.5336^	.4363*	.1539	.5868^	1.0000	–	–
T7	.2894	.3285	.1140	.2574	.1660	.3816	.4573*	.3086	.6348^	.2931	1.0000	–
CC	.5529^	.5871^	.4576*	.3284	.3059	.3062	.4869*	.2698	.2709	.2729	.2988	1.0000

N of cases: 42 2-tailed Signif: *– .01 ^ –.001

overall CC were collected on the basis of the total score of a series of tests given in a conversation course focusing on both aural comprehension and oral production skills.

Based on the dimensionality check which suggests that T3, T4, and T5 violate the unidimensionality assumption, IRT cannot be applied to TOEIC to estimate test-takers' ability. Thus, a comparison was made between CR4 and TOEIC using individual scores obtained by statistical procedures based on classical testing theory.

Given the fact that the CC was the total score of a series of measures of communicative competence, CC can be regarded as a valid meaure of global proficiency. The relatively low correlations between CC and TOEIC, and the relatively high correlations between CC and CR4 reveal that CR4 has higher construct validity than TOEIC. Among CR4 test batteries, the speeded tests — 4R and 4L — were highly correlated with CC. This indicates that the speeded test is a highly valid test type for measuring communicative competence, and that listening comprehension skills are highly correlated with overall communicative competence.

The low correlation between R4 and T7 demonstrates that the test content domains of the two reading comprehension tests diverge greatly from each other. CR4 covers non-esoteric academic reading passages as well as practical reading passages including business-oriented topics, whereas T7

focuses merely on business-oriented reading passages.

The relatively low correlation between T1 and other measures supports the hypothesis that the test method of picture description is concerned with constructs apart from what is measured by other tests. The unique factors may be heavily related to cultural awareness and/or exposure to motionless visual stimuli in addition to aural linguistic stimuli.

The overall low correlation between T4 and other measures suggests that exposure to the lengthy listening passage results in increased difficulty and lowered discriminability, and thus does not guarantee a high explanatory power of communicative competence. This finding suggests that the lengthy passage may not always enhance the validity of a test. It is worth noting that testing is inherently a sampling process of a test-taker's ability, especially so in language testing which deals with the extremely complex processes of language use. The validity of a test is enhanced only through maximized efficiency in sampling in terms of proper content domains and test methods along with appropriate difficulty including the length factor.

5.2. Interchangeability

As was mentioned in Dimensionality Check, it is not theoretically plausible to apply IRT-based true score equating due to the violation of the unidimensionality assumption by TOEIC data. Furthermore, the CR4 test batteries cannot be directly equated with TOEIC in that there is no one-to-one correspondence between the two tests in terms of the test method facets and the test content. The content of the CR4 reading passages (covering practical topics as well as non-esoteric academic topics) greatly diverges from that of TOEIC, which is concerned only with practical topics. Part V of TOEIC, which combines two content areas of vocabulary and grammar, cannot be compared with the vocabulary and the grammar of CR4. Therefore, as the content analysis suggests, the only part that can be equated lies in such aural comprehension areas as Parts II, III, and IV of TOEIC and CR4. As the comparability study does not offer much practical utility in terms of interpretation of total test scores, regression analyses were not conducted as was previously proposed.

6. Conclusions

6.1. Comparability between TOEIC and CR4

The test method facets and the content of CR4 and TOEIC were validated on the basis of the fundamental considerations of language testing. In the first place, there are two points to be made regarding the comparability between TOEIC and CR4. First, Stout's dimensionality check demonstrates that some of TOEIC test batteries fail to satisfy the assumption of essential unidimensionality. The overall disparity was found in the test content specifications and formats of the two test batteries. Thus, the violation of the dimensionality assumption and the mismatch in content domain invalidates the true score IRT equating, which was proposed to investigate the comparability between the two test batteries. Second, the correlational analyses reveal that CR4 serves as a more valid tool than TOEIC in measuring overall communicative competence. The second finding requires further empirical research focusing on the criterion-related validity or predictive utility with a larger sample of subjects.

6.2. Test Method Facets of TOEIC and CR4

In the second place, the test methods were investigated on the basis of item (distractor) analyses and language testing theories. First, the analyses recommend some of the revisions to be made in TOEIC test methods. Second, the analyses recommend that the test method facets of SNUCREPT remain as proposed on the basis of the analyses of the three CR research conducted during the three-year-long period ('91~'94). The findings of the present research validates the test characteristics of the SNUCREPT such as: 1) criterion-referenced interpretation of test results within IRT framework; 2) speeded test of grammar and vocabulary; 3) context-embedded (gap-filling) method in vocabulary and grammar; 4) spoken and written style in vocabulary and grammar; 5) OPOI (One-Passage-One-Item) principle in reading and listening tests; 6) balanced content domain of reading and listening passages; 7) appropriate test method for reading and listening content; 8) unadulterated listening test presented only in aural channel; and 9) being exposed to the listening passage twice.

References

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

Bachman, Lyle, Fred Davidson, Katherine Ryand, and Inn-Chull Choi (in press), *The Cambridge-TOEFL Comparability Study*, UK: Cambridge University Press.

Choi, Inn-Chull (1992a) 'The Interplay of Natural Phonology and Sociolinguistic Variable of Formality', *English Teaching* 44, 65-88, The College English Teachers Association of Korea.

Choi, Inn-Chull (1992b) *Theoretical Studies in Second Language Acquisition: An Application of Item Response Theory to Language Testing*, New York: Peter Lang Publishing Inc.

Choi, Inn-Chull (1994) 'Content and Construct Validation of a Criterion-Referenced English Proficiency Test', *English Teaching* 48, 311-348, The Korea Association of Teachers of English.

Harman, H. H. (1976) *Modern Factor Analysis*, (3rd ed.), Chicago: University of Chicago Press.

Henning, Grant and Eduardo Cascallar (1992) 'A Preliminary Study of the Nature of Communicative Competence', *TOEFL Research Reports* 36. Feb., Princeton: Educational Testing Service.

Oller, John W. Jr. (1979) *Languages Tests at School: A Pragmatic Approach*, London: Longman Inc.

Savignon, Sandra J. (1986) 'The Meaning of Communicative Competence in Relation to the TOEFL Program', In Charles W. Stansfield (Ed.), *Toward Communicative Competence Testing: Proceedings of the Second TOEFL Invitational Conference, Research Reports*, May, 1986, ETS.

Stout, William, R. Nandakumar, B. Junker, and H. H. Chang (1991) *Dimtest and Testsim*, Urbana, IL: University of Illinois.

Tatsuoka, Maurice M. (1988) *Multivariate Analysis: Techniques for Educational and Psychological Research*, (2nd Ed), NY: Macmillan Publishing Company.

Appendix A

1. Item # 53

A: How big a place are you interested in?

B: We'd like three bedrooms with two baths. And my husband insists having a garden.

A: Well, I'll see what I can do. With home values falling as they are, very few houses are going on the market right now.

Q: Where are they?

a. At a real estate office

b. In a hotel lobby

c. At a market

d. In a three-bedroom house

2. Item # 61

A: Pardon me, sir. The camera has to go through too. The fill will be unaffected.

B: But, this is extremly light-sensitive film.

A: All right. Put it on this tray and pass through the metal detector please.

Q: Where are they?

a. At a camera shop

b. In a print lab

c. At a security checkpoint

d. In a photography class

3. Item # 74

A: Can you give me change for a one-dollar-bill?

B: I'm sorry, but bus drivers aren't allowed to make change.

A: Well, just take it then. I sure don't have enough to take a taxi.

Q: Where are they?

a. In a bank

b. On the subway

c. In a taxi

d. On a bus

4. Item # 80

A: How are you two doing over here? Let me get you some fresh drinks.

B: Just coffee for me. I'm the driver tonight.

A: Glad to hear it. I suppose you heard Fred got stopped by the police for driving under the influence after our last get-together.

Q: Where are they?

a. In a drive-in

b. At a party

c. At a coffee shop

d. In a police station

5. Item # 94

(See Item # 94 under 3.5.)

6-7. Questions 193-195.

Dosage: Adults, 2 tables every 6 hours, not to exceed 8 tablets in 24 hours, or as directed by physician. Children under 12 should use only as directed by physician.

Warning: Do not exceed recommended dosage. If symptoms persist, do not improve within 7 days, or are accompanied by high fever, or if new symptoms occur, see your doctor before continuing use. Do not take this product if you have high blood pressure, heart disease, diabetes, thyroid disease, or difficulty in urinating, except under doctor's supervision. Do not take this product for more than 10 days.

6. Item # 193

What is the purpose of this notice?

a. To advertise a medicine

b. To give instructions for the taking of a medicine

c. To warn people not to take a certain medicine

d. To recommend visiting a doctor

7. Item # 195

What should the patient do if he develops a high fever?

a. He should take the medicine to reduce his fever.

b. He should increase the dosage.
c. He should stop taking the tablets.
d. He should take the medicine for 10 days.

Appendix B

Eigenvalues from the Factor Analysis in Different Tests

(For the explanation of how to interpret eigenvalues in factor analysis, see Harman, 1976, and Tatsuoka, 1988).

1. 4L

Factor	Eigenvalue	Difference
1	9.95409	6.19779
2	3.75630	.21997
3	3.53633	.37818
4	3.15815	.30313
5	2.85502	.09476
6	2.76026	.07274
7	2.68752	.16202
8	2.52549	.46913
9	2.05637	.12233
10	1.93404	.09828
11	1.83576	.13281
12	1.70295	.16544
13	1.53750	.15164
14	1.38587	.06749
15	1.31838	.03032

2. 4G

Factor	Eigenvalue	Difference
1	7.45895	.70300
2	6.75595	3.29234
3	3.46361	.33054
4	3.13308	.09049
5	3.04259	.64536
6	2.39723	.09226
7	2.30496	.16639
8	2.13857	.20247
9	1.93610	.09088
10	1.84521	.15792
11	1.68730	.21941
12	1.46789	.33899
13	1.12890	.04739
14	1.08150	.09569
15	.98581	.02789

3. 4V

Factor	Eigenvalue	Difference
1	16.59385	12.10622
2	4.48763	.85304
3	3.63459	.78338
4	2.85121	.08336
5	2.76785	.64452
6	2.12333	.09696
7	2.02637	.14879
8	1.87758	.18637
9	1.69121	.26099
10	1.43022	.12672
11	1.30350	.10230
12	1.20120	.11125
13	1.08995	.09046
14	.99949	.05087

4. 4R

Factor	Eigenvalue	Difference
1	6.04795	2.01920
2	4.02874	.79736
3	3.23138	.74233
4	2.48905	.42408
5	2.06497	.34337
6	1.72160	.25532
7	1.46628	.26563
8	1.20065	.05318
9	1.14747	.11520
10	1.03227	.12221
11	.91005	.08104

5. T1

Factor	Eigenvalue	Difference
1	2.73076	1.00808
2	1.72268	.38705
3	1.33562	.16633
4	1.16929	.17787
5	.99142	.17568

6. T2

Factor	Eigenvalue	Difference
1	8.40659	4.96937
2	3.43721	1.05670
3	2.38052	.09470
4	2.28582	.28668
5	1.99914	.05610
6	1.94304	.33280
7	1.61024	.15431
8	1.45593	.10231
9	1.35362	.06479
10	1.28883	.12353
11	1.16530	.15394
12	1.01136	.05789
13	.95347	.03468

7. T3

Factor	Eigenvalue	Difference
1	6.93369	3.84363
2	3.09006	.29822
3	2.79185	.21231
4	2.57953	.42772
5	2.15181	.44884
6	1.70297	.32730
7	1.37566	.28028
8	1.09538	.05236
9	1.04302	.05942
10	.98360	.18615

8. T4

Factor	Eigenvalue	Difference
1	2.09323	.24012
2	1.85311	.41349
3	1.43962	.19662
4	1.24301	.21829
5	1.02472	.03726
6	.98746	.22274

9. T5

Factor	Eigenvalue	Difference
1	7.52093	4.00519
2	3.51574	.90125
3	2.61449	.28887
4	2.32562	.12365
5	2.20197	.12722
6	2.07475	.17590
7	1.89884	.25842
8	1.64042	.19534
9	1.44508	.21740
10	1.22768	.08028
11	1.14740	.09141
12	1.05599	.10551
13	.95048	.01612

10. T6

Factor	Eigenvalue	Difference
1	3.61923	2.18838
2	1.43084	.13539
3	1.29546	.16623
4	1.12923	.23251
5	.89672	.02793

11. T7

Factor	Eigenvalue	Difference
1	7.45916	4.06265
2	3.39651	.65711
3	2.73940	.28329
4	2.45611	.24289
5	2.21322	.27655
6	1.93667	.21324
7	1.72343	.10413
8	1.61931	.32255
9	1.29676	.16449
10	1.13226	.05306
11	1.07920	.11148
12	.96772	.12707

Department of English
Sungshin Women's University
249-1, Dongseondong-3Ka,
Seongbuk-Ku, Seoul
136-742
Korea