9月16日(土) 小講演第4室（733）

Content and Construct Validation of a Criterion-Referenced English Proficiency Test

Inn-Chull Choi
(Dept. of English, Sungshin Women's University)
(Language Research Institute, Seoul Nat. Univ.)

I. INTRODUCTION

In globally competitive modern societies, it is increasingly important to develop a valid tool by which to measure language proficiency from the perspective of criterion-reference (C-R) as well as norm-reference (N-R). TOEIC (Test of English for International Communication) and TOEFL (Test of English as a Foreign Language) have been considered as a valid tool to meet such a demand. TOEIC and TOEFL are said to measure the overall proficiency required for international business and for academic work in English-speaking setting, respectively. Some questions were raised, however, as to the validity of the test method facets of the two tests on the basis of the recent research on the development a standardized EFL test battery entitled Seoul National University Criterion-Referenced English Proficiency Test (SNUCREPT) (Choi, 1994).

II. RESEARCH METHODS

Vadid data were obtained from approximately 1,000 students from five different universities. 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.

The whole process for construct validation involved qualitative and quantitative approaches pertaining to each issue in question, e.g., 1) need for systematic test development, 2) analyses of descriptive statistics and test/item indices (including reliability, facility, and discriminability), 3) item and distractor analyses, 4) quantitative and qualitative approaches to test method analyses, 5) content and TMF-based difficulty analyses, 6) IRT application for more precise measurement, and 7) correlational analyses. The findings lay a fundamental basis on which to determine the extent to which TOEFL, TOEIC, and SNUCREPT are valid and comparable to each other.

III. ANALYSES of DESCRITIVE STATISTICS

The descriptive statistics show that the systematic test development results in relatively high reliability indices over .82. The mean facility indices are approximately .5 across the test batteries. The mean discriminability indices for all the tests are over .35, which show the adequacy of the power of discrimination of SNUCREPT.

The distractor analyses reveal that the nature of passage topics constitutes one of the most significant factors determining difficulty of reading comprehension tests. This is the very point in which the dilemma lies in accommodating construct validity. Granting the fact that the reading process is heavily influenced by the individual reader's background knowledge or schemata, the reading comprehension test should be far from a measure of background knowledge related to the reading passage, unless the test is intended particularly for ESP. Therefore, this reading/listening passage factor should always be given the utmost consideration to minimize the effect of test-takers' background knowledge on their test performance.

IV. TEST METHOD FACETS

As for the major factors influencing test results, many studies have indicated that test results are influenced not merely by test-takers' ability but also by test methods (Bachman and Palmer, 1982; Shohamy, 1983, 1984; Bachman et al, 1995). Thus, in order to ensure maximal validity, it is necessary to investigate the extent to which the given test methods are valid. The most noteworthy findings based on the qualitative and quantitative analyses to verify the validity of test formats are the followings.

1. Listening Comprehension

A. *Unadulterated Listening Task Only*

The most salient feature of SNUCREPT listening test is the 'Aural Only' mode. It is aimed at measuring aural comprehension only. Other tests like TOEFL or TOEIC are designed in such a way that the test-takers are required to *read* (not listen to) the choices after listening to the question. Thus, they are likely to obtain listening test results seriously adulterated with reading ability.

It should also be pointed out that in the case of TOEFL or TOEIC it is impossible to control the test-takers regarding whether they read the choice before listening to the listening passage or vice versa, or read and listen simultaneously. It is evident that the test results vary significantly depending on what kind of the test-taking strategy the test-takers choose to employ.

It is also worth observing that in the case of TOEFL or TOEIC while the directions for each part are being read, the test-takers with 'test-wiseness' prepare for the test by reading the choices. As instructed by the directions, on the other hand, some naive test-takers read and listen to the instructions simultaneously. It is certain that this seemingly innocuous format does function as a significant test bias factor.

B. *Two Time Exposure to the Aural Passage*

Like the Cambridge Proficiency Test Batteries of the University of Cambridge Local Examination Syndicate (UCLES), SNUCREPT Listening Test allows test-takers to listen to the questions twice and to the articulately presented choices once. This test method proves valid especially in a Question and Answer format. The test-takers listen to the passage first in a macro-listening manner, and then to the question. Then they understand what they are expected to listen for. This enables them to listen for specific information when they listen to the passage again in a micro-listening way. This issue can be easily solved only if we can utilize the video facility for listening tests in which test-takers are provided with all the visual clues describing the circumstances in motion pictures.

C. *One-Passage-One-Item (OPOI) Principle*

The OPOI principle is essential to local independence, one of the fundamental assumptions of IRT. Two or more items relating to one listening passage are likely to be locally dependent, i.e., heavily associated with each other in terms of their functions with the test-takers. The concept of local dependence is seriously detrimental to test objectivity or fairness, especially when the passage deals with technical topics or jargon. Factor analysis clearly demonstrates that the topic factor is one of the most salient factors in such a test format. Hence, it is imperative that the OPOI principle be employed throughout the test. It has been discovered that lengthy listening passages as in TOEFL Listening Section C burden even native speakers with a heavy memory load.

2. **Grammar**

A. *Explicit Knowledge of Grammar: Context-Reduced Error-Detection Task*

The TOEFL Section 2 structure test format, in which four parts are underlined and the erroneous part is identified. This format reflects the characteristic of discrete-point test (i.e., low reliability), which induces 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, 1982).

A more desirable test format is to divide a sentence into four parts with three slashes. It is valid in that test-takers are expected to deal with 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 more valid method is to provide a more context-embedded discourse or dialogue.

B. *Time Allocation*

The grammar test is designed to validly measure the test-taker's acquistion, not learning, through restraining the use of monitor (Krashen, 1985) or test-taking 'wiseness' strategy. It is to approximate a speed test by maximizing the speediness of the test (Inn-Chull Choi, 1991). A survey of Korean test-takers indicates that the 25-minute-long TOEFL grammar test is far from a speed test for the majority of test-takers. Thus, TOEFL grammar test is considered inadequate to measure the authentic grammatical competence or the acquistion (Krashen, 1985).

3. **Vocabulary**: *Synonym of Underlined Part*

This task has the lowest reliability (.392) index, which is quite understandable considering the process which the test-takers go through in dealing with this test method. As the test-takers are not forced to read the entire stem to answer, they are likely to focus on the word without referring to the context. This format leads to low reliability and consequently low validity. TOEFL or TOEIC Vocabulary test uses this typical test method, which is of the discrete-point test format motivating the test-takers to think little of context. This problematic method should be replaced by gap-filling task, which proves to be fairly reliable, whether spoken or written English.

4. **Reading**

A. *Topic Factor & One-Passage-One-Item (OPOI) Principle*

As in the SNUCREPT Listening test, the OPOI Principle is applied to the SNUCREPT reading test, which is pivotal to local independence. This priciple enables us to exclude the topic factor, which has been shown to play a predominant role in causing test-bias. TOEFL Reading test has the typical problem of topic factor bias (Choi, 1991), in that It has about five to six academic/technical reading passages, each of which is followed by four to six question items. Those test-takers with adequate background knowledge in the relevant field will find the reading question items very easy to solve, and vice versa. Granting the fact that the background knowledge is fundamental to successful reading, a valid language test should minimize the influence of this factor on the test performance.

B. *Gap-Filling Task vs. Question-and-Answer Task*

The gap-filling task proves to be very reliable. It is probably attributed to the fact that the method is inherently free from much of the bias caused by developing the choices, especially the key. In the multiple-choice format, the validity of a test is literally a function of the quality of the keys and distractors invented by the test writers. Unlike other multiple choice formats, this format does not require the test writer to develop his own key since it is already provided in the original text. The fact that original text is used for developing keys minimizes the artificiality of inventing the text of keys (Choi, 1991).

V. PREDICTABILITY of DIFFICULTY

The most single salient index of language testing is the concept of difficulty, which constitutes the pivotal component of validity. Being a function of a wide range of variables, such as Test Method Facets, task types, and most importantly, the test-takers' ability, the difficulty level of a test is extremely difficult to predict prior to the administration of the test. To maximize the validity and reliability of a test, it is crucial to be able to predict the difficulty level on the basis of Test Method Facets. To establish a procedure for predicting the difficulty level, we need to determine the potential variables known to affect the readability, and consequently, the difficulty. Bachman's TMF is employed to investigate the extent to which variables influence the difficulty level in a most significant way (Bachman, 1990). Due to the limited availability of robust instruments, the present study explores nine instruments used to operationalize the above variables as follows:

First, the length facet is represented by (1) total number of words, (2) total number of clauses, (3) total number of sentences, (4) average number of syllables per word, (5) average number of words per sentence. Secondly, the propositional content facet consists of the type of information measured by (6) total number of abstract words divided by total number of content words, and the distribution of information measured by (7) total number of content words divided by total number of words, and the topic measured by (8) technicality and culture-specificity. Thirdly, the organizational characteristics is reflected by (9) the total number of clauses divided by total number of sentences. Fourthly, two more indices are used to indicate the degree of readability, i.e., 100 minus Flesh index, and Fog index. Two indices are utilized to reflect the extent of difficulty level: (10) IRT difficulty index b and (11) 1 minus facility index (P). Finally, (11) difficulty indices predicted by the test developers are presented.

The correlation analyses of the above variables reveal reveal the following findings. First, the correlations between the two readability indices and the two difficulty indices are so low that they are almost negligible. The high correlation (.8788) between the two readability indices, however, show that they are closely interrelated with each other. It is noteworthy that the two indices have very high correlations with SW). This suggests that these indices are based primarily on mechanical aspects of text, such as average number of syllables per word. Second, the total number of clauses divided by the total number of sentences (CS) and the total number of abstract words divided by total number of content words (AC) have statistically significant correlations with B (IRT Difficulty Index: .4002, .5376) or -P (1 - Facility Index P: .4154, 6186). This finding indicates a possibility that CS and AC will serve as solid indicators of readability. The total number of abstract words divided by the total number of content words can be construed as a degree of abstraction of the total amount of information to be processed.

VI. IRT-BASED PRECISE MEASUREMENT

1. More Precise Test/Item Statistics

The test statistics include three item indices, i.e., a: discrimination; b: difficulty; c: guessing. Compared with classical testing theory (CTT), IRT's probabilistic estimation makes it possible for us to obtain a more precise measurement of difficulty and discrimination than the classical testing theory which is based on the facility index of proportion correct (p) and of discriminability index r_{bis} (D). The more precise measurement tool enables us to identify the behavior of each item and ultimately to develop a more valid C-R test.

2. More Precise Test-taker Statistics

The ability index is denoted by Θ, which is estimated along with the item indices, a and b. The probabilistic model of IRT allows us to estimate more precise measurement of ability in a criterion-referenced fashion than does CTT which depends merely on the number correct.

In the case of the reading test, the results in the following table illustrate that two test-takers who had the same number correct of 35 (out of 50) with CTT tools, were found to have different ability levels of .8974 and 1.1340 within IRT framework. This discrepancy between CTT and IRT is based on the fact that IRT estimates individual ability level through simultaneous consideration of difficulty and discriminability of each item, which highlights the most significant superiority of IRT over CTT.

Sample Ability Indices of Reading Comprehensione

	Content	#	#	CTT	IRT
ID #	Component	Tried	Right	(%)	(Θ)
XXXXXXXX	TOTAL	50	35	.7000 ¦	.8974
XXXXXXXX	TOTAL	50	35	.7000 ¦	1.1340

It is hoped that the present research sheds some light on how to apply the empirical evidence and the theoretically sound frameworks to the overall processes of effective test development and rigorous construct validation in language testing.

VII. REFERENCES

Bachman, Lyle F. (1986). The test of English as a foreign language as a measure of communicative competence. In C. W. Stansfield (Ed.), Toward communicative competence testing: proceedings of the 2nd TOEFL invitational conference. *Research Reports.* May, 1986. ETS.

Bachman, Lyle F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, Lyle F. and Palmer Adrian. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly,* 16, 4, 449-65.

Bachman, Lyle F., Fred Davidson, Kathy Ryan, and Inn-Chull Choi. (1995). Studies in Language Testing 1: An investigation into the comparability of two tests of English as a Foreign Language. United Kingdom: Cambridge University Press.

Canale, Michael. (1983). On some dimensions of language proficiency. In John W. Oller, Jr. (Ed.), *Issues in language testing research.* Rowley, MA: Newbury House. 333-42.

Choi, Inn-Chull. (1989). *Past, present, and future of language testing.* English Teaching. 38, 95-135. The College English Teachers Association of Korea.

Choi, Inn-Chull. (1991). *Theoretical studies in 2nd language acquisition: Application of Item Response Theory to language testing.* New York: Peter Lang Publishing Inc.

Cziko, Gary. (1983). Psychometric and edumetric approaches to language testing. In John W. Oller, Jr. (Ed.), *Issues in language testing research.* Rowley, MA: Newbury House. 289-308.

Krashen, Stephen D. (1985). *The input hypothesis.* London: Longman Inc.

Oller, John W. Jr. (1979). *Languages tests at school: a pragmatic approach.* London: Longman.

Richards, J. C. (1985). *The context of language teaching.* NY: Cambridge Univ. Press.

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 2nd TOEFL invitational conference. *Research Reports.* May, 1986. ETS.

Shohamy, Elana. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing,* 1, 2, 147-70.

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