Construct Validation Study on SNUCREPT (Seoul National University Criterion-Referenced English Proficiency Test)*

Inn-Chull Choi

The first pilot test for Seoul National University Criterion-Referenced English Proficiency Test (SNUCREPT) was developed to conduct a validation study to set a framework on which to develop the standardized EFL test battery, which will be used to measure English proficiency in a criterion-referenced manner. Designed to be a speed test, the total test of 250 items lasted approximately 150 minutes. Two analytic approaches, quantitative and qualitative approaches were employed to conduct the content analyses.

Approximately 1,000 college students participated in this study. The descriptive statistics show that the distribution approximated normal distribution. The test indices (reliability, facility, discriminability) meet the requirement of a desirable CR test. The test methods were analyzed qualitatively and quantitatively to investigate the extent of their authenticity. Some of the findings help identify the problematic formats of widely used EFL test batteries, thus providing invaluable information for improving the formats of SNUCREPT. Application of Item Response Theory (IRT) was validated through checking the dimensionality assumption with Stout's factor analysis methods. IRT allows us to estimate more precise measurement of ability and item indices than does classical testing theory. Correlational analyses were done to validate the SNUCREPT in relation to Test of Spoken English (TSE). The results show a high correlation between TSE and SNUCREPT, which supports the rationale of language tests being speed tests.

The first pilot test results lay a systematic basis on which to revise the format of SNUCREPT and to develop the second pilot test, which is to be further validated to finalize the testing specifications of SNUCREPT.

* This paper is an interim report on the SNUCREPT development project funded by SNU LRI. My heartfelt appreciation goes to Profs. Nahm-Sheik Park, Choong-Bae Kim, Dae-Sik Min, Ki-Sun Hong, Kyoung-Goo Shin, Byoung-Gyu Ahn, Hee-Rahk Chae, Young-Sook Lee, Jeong-Mi Yoon, Robert Fouser and SNU Language Research Institute staff members for providing me the data used in this study.

I. Development of 1st Pilot Test for SNUCREPT

1. Rationale

The 1st pilot SNUCREPT was developed and analyzed to conduct a construct validation study on test content and format. The results were used to revise the content and format of the 1st pilot test and subsequently to be used as the guideline for developing the 2nd pilot test. The final analysis of the 2nd pilot test results will lay the foundation upon which to establish the standardized guidelines of Seoul National University Criterion-Referenced English Proficiency Test (SNUCREPT, for short tentatively) intended to measure overall English proficiency of the SNU students and the larger public.

2. Test Characteristics

2.1. Criterion-Referenced (C-R) Test

The test is designed to maximize the features of a C-R test, i.e., well-defined content domain and clearly-specified criterion/grade, which are extremely difficult to operationalize. The former feature is based on Bachman's (1990) framework of communicative language ability (CLA). The latter is facilitated by IRT-driven precise measurement (Choi, 1991).

2.2. Time Restriction (Speed Test)

The most significant feature of the SNUCREPT lies in its time limitation, especially in grammar and vocabulary tests. Supported by the notion of acquisition vs. learning (Krashen's Monitor Model, 1985), the test is designed to maximize speediness in order to measure the acquired (subconscious) communicative competence, while inactivating the conscious learning. As far as real-time communicative setting (e.g., listening to genuine broadcasts) is concerned, conscious-level learning does not help improve aural comprehension. The consciously learned system even prevents the listeners with high intolerance of ambiguity from skipping the missed aural input and resorting to prediction which is most crucial to successful comprehension. Forcing them to dwell too long on the unintelligible input, their

learning system causes the listeners to fall behind the seemingly fleeting speech in real-time communication. Consequently, the learning (unless internalized and automatized) may well be detrimental to aural comprehension of normal speed speech (Choi, 1988).

The factor of time allocation for the test is determined on the basis of the three native speakers' performance on the pilot test. The average performance in test-taking time is provided in parentheses in the following table.

2.3. Content Domain & Rubric

The structure, the time allocation and the scoring weights of 1st pilot test are as follows:

Table of Specifications for 1st Pilot Test

Test	# Items	Length (min)	Total Score (pts)
Listening:	100	60	400
Grammar:	50	15 (7)	100
Vocabulary:	50	15 (6)	100
Reading:	50	60 (26)	400
Total:	250	150	1000 Max Score

NB: ( ) under length denotes average performance of native speakers

3. Content Analyses

The test content in terms of readability or difficulty was analyzed on the basis of Bachman's (1990) framework which incorporates the current theoretical models of language ability (Canale, 1983; Cummins, 1983). The methodology included qualitative and quantitative approaches. The former utilized linguists' judgment and native speakers' insight. The latter was based on the analysis done by computational linguistic tools.

To maximize reliability and validity through minimal test bias, characteristics of the integrative test (for measuring overall proficiency; Oller, 1979; Cziko, 1983) are fully employed and a variety of TMF aspects are systematically analyzed for the development of the pilot test. The currently available test formats which are considered to be relatively desirable are chosen to be validated. The major considerations for the TMF of each test are as follows:

3.1. Listening

For the content of the listening test, taken into consideration are the four major factors: subskill components, test methods, content (functions and notions), difficulty levels.

On the basis of the theoretical model of listening skill components or strategies (Rivers and Temperley, 1978; Ur, 1984; Richards, 1985), the subskill components to be validated focus on identifying 1) segmentals/ suprasegmentals, 2) lexicons and syntactic structures for basic communicative fuctions, 3) grasping detailed information, 4) understanding gist/topic, 5) logical inference, and 6) inferring tone/attitude. The range of test methods to be validated include 1) describing a picture (simple description and question/answer type), 2) identifying an identical sentence, 3) paraphrasing, 4) inferring an appropriate response (based on single sentence and dialogue), 5) question and answer (based on short dialogue, discourse, and long dialogue). Content topic is divided into 20 function/notion/situation categories such as greeting, apologizing, thanking, warning, etc. The difficulty of the test questions range is divided into five levels.

For foreign language tests, the level of difficulty is principally a function of speech speed and the consequent phonological variations, which depend on the formality of context (Choi, 1992). These factors should be carefully considered in developing valid listening tests (especially in the actual recording setting), which will have the most significant washback effects on teaching listening comprehension and language teaching in general.

3.2. Grammar

For the grammar test, among the four factors are the contextuality of stems, test methods, topic (formality), difficulty. The pivotal aspect of the grammar test is the degree to which the stem is context-embedded. Every effort should be made to maximize the contextual-embeddedness. The test methods to be explored include 1) gap-filling (based on single sentence in written language), 2) gap-filling (based on dialogue in spoken language), 3) detecting grammatical errors (of segments grouped in a sentence), 4) choosing the grammatically correct sentence, and 5) choosing the grammatically incorrect sentence. Based on the findings of contrastive analysis or interlanguage research, the grammatical points are believed to have an in-

herent degree of difficulty, which requires further research. The difficulty range is divided into five levels.

3.3. Vocabulary

Like the grammar test, the vocabulary test is based on four factors such as the contexuality of stems, test methods, topic (formality), difficulty. The most crucial aspect of the vocabulary test is the extent to which the stem is context-embedded (Madsen, 1983). Every effort is made to maximize contextual-embeddedness. The test methods to be investigated include 1) paraphrasing, 2) identifying synonym for underlined part, 3) gap-filling (based on dialogue in spoken language) and 4) gap-filling (based on single sentence in written language). The topic is related to the degree of formality of the stem, which in turn is closely associated with the notion of dichotomous presentation format, i.e., spoken or written langauge. The formal topic areas in single sentence format follow those of the reading test, whereas the informal topic areas in dialogue format are similar to those of the listening test. The difficulty, which is mainly based on frequency and semantic network, is divided into five levels.

3.4. Reading

For the reading test, the four major factors are considered: subskill components, test methods, content (reading passage topics), and difficulty levels.

Based on the theoretical models of reading skill components or strategies (Goodman, 1967; Smith, 1982; Dubin, Eskey and Grabe, 1986), the subskill components to be validated include 1) identifying detail information, 2) grasping gist/topic, 3) inferring contextual logic, 4) understanding coherence/cohesion and 5) inferring tone/attitude. The variety of test methods to be validated focus on 1) question and answer type (based on reading a passage for detail information, gist/topic, inference), 2) gap-filling, 3) reorder (for coherence) 4) insertion (for coherence), 5) deletion (for coherence). The test is separated into two parts in terms of topic, i.e., academic-related passages (written in plain language) and non-academic. Each section is categorized into four subsections. The difficulty range is divided into five levels.

3.5. Speaking & Writing

These productive skills are to be tested in the production mode for valid measurement. The limited availability of logistices makes it difficult to implement tests with valid methods. The practicality of rating poses another serious problem in measuring the productive skills. For large-scale administration of standardized tests, analytical (e.g. as in TWE, TSE) as well as integrative or global (e.g. as in Cambridge Proficiency Test batteries) rating systems should be explored. Further research is required for the development of valid and practical test methods and rating procedures.

4. Test Writers

Six applied and theoretical linguists were test writers for the 1st pilot test. This test was then proofread by two native speakers of English with ESL background. The test writers were trained to take the aforementioned four factors or dimensions into account simultaneously in developing the test and analyzing its content. It should be noted that this notion of dimension is far from the two dimensional model of Lado's skill-by-component framework (1961).

II. Descriptive Statistics*

1. Samples

Over 700 students from five different universities (including Seoul National Univ., Korea Univ., Konkuk Univ., Chunnam National Univ., and Kwangju Univ.) participated in taking the test. The consideration of test security also 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 represent a wide range of stu-

* As the test is designed to reflect the qualities not only of C-R and N-R, but also of speed and power, the conventional statistics are presented hereafter. It should be noted that C-R and N-R, speed and power are the concepts on the continuum.

dent 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.

2. Data Description

The descriptive statistics are as follows:

SNUCREPT listening, grammar, vocabulary, reading test will be denoted by CL, CG, CV, CR, respectively hereafter.

Descriptive Statistics

	CL	CG	CV	CR
N of Items	100	50	50	50
N of Examinees	738	1127	1106	1119
Mean	44.305	23.443	24.396	29.449
Variance	173.057	48.007	42.483	53.697
Std. Dev.	13.155	6.929	6.518	7.328
Skewness	0.612	−0.016	0.051	−0.398
Kurtosis	0.089	−0.354	0.098	−0.017
Minimum	10	4	3	4
Maximum	92	44	43	45
Median	42	24	24	30

2.1. Mean

The statistics show that the mean is almost negligibly less than .5., which means that the overall proportion correct across the test batteries is approximately .5.

2.2. Standard Deviation

The relatively great standard deviations are obtained all across the test batteries. This reasonably wide dispersion is desirable for a test intended to discriminate the test-takers' performance.

2.3. Skewness

The skewness index represents the extent to which a test is symmetric.

The index for CL is shown to be slightly positively-skewed, which indicates that the test's overall difficulty is slightly difficult. The index for CR is proven to be slightly negatively-skewed, which suggests that the test as a whole is almost negligibly easy. The indices for CG and CV turn out to be near zero. The minimum and maximum scores also demonstrate the fact that perfect scores are not obtained across the tests, whereas the three native speakers got perfect scores on all the tests. This finding should be considered in adjusting the difficulty level of the 2nd pilot test.

2.4. Kurtosis

As was indicated by the standard deviation, the kurtosis or peakedness indices suggest that all the tests are near zero, which reflect the shape of normal distribution. Only the fact that the index for CG is slightly negative is considered desirable in that the more negative kurtosis the more spread-out the distribution. This is desirable for maximizing the discriminability.

2.5. Implication

All the indices indicate that the overall data do not seriously violate the assumptions of normal distribution. This ensures the legitimate employment of the following statistical analyses.

3. Test/Item Indices

The systematic test development results in relatively high reliability indices over .82. The mean facility indices are approximately .5 across the tests. The interpretation of the difficulty level is mentioned above. The mean discriminability indices for all the tests are over .35, which show the adequacy of the power of discrimination of SNUCREPT (Gronlund, 1985).

Test Indices

	CL	CG	CV	CR
Reliability (alpha)	0.857	0.854	0.825	0.861
Std Estimate of Measurement	4.466	2.956	2.987	3.016
Mean Facility (p)	0.443	0.469	0.488	0.589
Mean Discriminability (r_{bis})	0.366	0.421	0.412	0.444

4. Item & Distractor Analyses

Based on the analyses of the facility and discriminability indices of all the test items, the problematic items are investigated in test content and format. Divided into the three ability levels, the data are reinvestigated to explore the function between item indices and test-takers' behavior. Moreover, the content of keys and distractors of the problematic items are analyzed to identify potential problems.

The following is the table of facility indices sorted in descending order. The table demonstrates that the facility indices range from near 0 up to over .9, which corresponds to the original design. Another point worth noting is that the table shows that results roughly correspond to the original design of item facility in terms of the distribution in descending order.

4.1. Facility Index

of Items across Facility Index Groups

Range	CL	CG	CV	CR
$.0 < p \leq .1$	1	1	2	3
$.1 < p \leq .2$	5	5	4	6
$.2 < p \leq .3$	9	10	5	9
$.3 < p \leq .4$	18	6	8	9
$.4 < p \leq .5$	10	1	3	4
$.5 < p \leq .6$	16	6	7	9
$.6 < p \leq .7$	17	8	7	4
$.7 < p \leq .8$	14	8	8	4
$.8 < p \leq .9$	4	3	5	1
$.9 < p \leq 1.$	6	2	1	1

Facility Indices
(sorted in descending order)

#	CL	#	CG	#	CV	#	CR
7	0.983	10	0.907	4	0.941	26	0.929
3	0.974	6	0.887	3	0.927	13	0.929
6	0.922	7	0.879	31	0.868	16	0.911
68	0.906	22	0.851	8	0.856	19	0.880
42	0.905	5	0.848	6	0.827	1	0.874
5	0.905	14	0.835	5	0.800	28	0.864
2	0.836	1	0.794	34	0.797	12	0.856

#	CL	#	CG	#	CV	#	CR
40	0.828	8	0.781	14	0.749	25	0.853
23	0.828	18	0.776	19	0.713	24	0.801
13	0.810	9	0.765	2	0.711	4	0.798
9	0.784	2	0.765	21	0.704	5	0.793
4	0.784	13	0.758	35	0.699	9	0.793
85	0.769	32	0.735	15	0.683	29	0.770
64	0.769	4	0.722	11	0.677	36	0.741
88	0.769	23	0.714	17	0.663	6	0.736
25	0.767	3	0.709	7	0.651	33	0.730
95	0.761	12	0.693	25	0.615	18	0.723
46	0.759	31	0.691	20	0.613	17	0.723
31	0.759	16	0.657	27	0.610	15	0.691
1	0.750	24	0.639	28	0.585	37	0.683
77	0.744	36	0.608	9	0.567	41	0.678
18	0.733	15	0.601	29	0.515	23	0.678
67	0.718	39	0.523	16	0.487	20	0.670
33	0.707	20	0.497	43	0.485	14	0.657
56	0.684	19	0.490	12	0.485	39	0.634
28	0.681	27	0.474	26	0.485	8	0.626
20	0.672	37	0.454	24	0.469	27	0.620
19	0.664	35	0.418	30	0.458	31	0.550
72	0.658	34	0.415	32	0.442	38	0.531
48	0.655	33	0.387	46	0.383	7	0.518
59	0.650	11	0.387	44	0.378	35	0.505
30	0.647	21	0.379	42	0.349	11	0.492
93	0.641	41	0.376	10	0.342	30	0.479
65	0.641	25	0.356	37	0.342	44	0.471
55	0.641	17	0.345	22	0.335	40	0.463
24	0.638	44	0.330	23	0.312	32	0.445
22	0.638	46	0.302	40	0.292	46	0.437
47	0.629	45	0.296	1	0.289	34	0.416
61	0.624	48	0.286	33	0.264	49	0.406
43	0.621	38	0.284	48	0.239	45	0.403
57	0.607	28	0.265	36	0.235	2	0.398
53	0.598	26	0.258	49	0.226	3	0.385
10	0.595	42	0.237	41	0.223	22	0.374
38	0.586	40	0.216	38	0.216	50	0.306
75	0.581	47	0.211	18	0.173	42	0.262
14	0.578	49	0.173	39	0.162	48	0.246
16	0.560	50	0.168	47	0.137	43	0.243
15	0.543	43	0.162	50	0.123	47	0.236
35	0.526	30	0.095	45	0.105	10	0.199
92	0.521	29	0.021	13	0.077	21	0.097

#	CL	#	CG	#	CV	#	CR
34	0.517						
73	0.513						
39	0.509						
74	0.504						
82	0.504						
96	0.504						
21	0.500						
78	0.496						
29	0.491						
80	0.487						
8	0.483						
11	0.483						
17	0.440						
52	0.427						
97	0.419						
60	0.419						
27	0.405						
86	0.393						
70	0.385						
50	0.379						
99	0.376						
44	0.371						
58	0.368						
83	0.359						
41	0.353						
51	0.350						
87	0.325						
91	0.325						
62	0.325						
36	0.319						
37	0.319						
26	0.319						
49	0.310						
100	0.308						
12	0.302						
45	0.284						
71	0.265						
90	0.256						
81	0.256						
54	0.248						
66	0.239						
89	0.239						
76	0.214						

#	CL	#	CG	#	CV	#	CR
79	0.205						
98	0.197						
84	0.188						
32	0.172						
63	0.162						
94	0.128						
69	0.094						

4.2. Discriminability Index

The following is the table of discrimination indices sorted in descending order. The table demonstrates that the discrimination indices range from near 0 up to over .9. The content analysis of the items with negative discriminability shows that all the items deal with the most important English content domain. The negative index is obtained when even the test-takers with high-level ability resort to wild guessing. Given this fact, this finding clearly reveals that most of the Korean college students are denied opportunities to learn about the essential content domain, especially in vocabulary and grammar.

The items with negative indices account for less than 5 percent of the entire test. It is thus determined that some of those items with the valid content should remain in order to exert a positive washback effect on the English education. For the listening test, however, the items with negative indices are shown to have rather lengthy content. This finding strongly suggests that a longer listening text does not help improve discriminability of aural comprehension on the part of Korean college students.

SNU CREPT Discriminability Indices
(sorted in descending order)

#	CL	#	CG	#	CV	#	CR
88	0.634	10	0.804	31	0.785	26	0.909
69	0.633	32	0.687	34	0.667	13	0.706
42	0.603	39	0.676	3	0.587	39	0.611
58	0.555	36	0.620	35	0.581	24	0.610
44	0.555	18	0.594	5	0.547	19	0.592
9	0.528	13	0.586	44	0.544	36	0.583
92	0.491	41	0.572	14	0.539	41	0.574

#	CL	#	CG	#	CV	#	CR
52	0.473	5	0.568	27	0.536	38	0.522
93	0.472	7	0.566	46	0.529	44	0.521
77	0.472	1	0.507	43	0.497	37	0.481
82	0.456	12	0.502	17	0.492	29	0.474
64	0.455	48	0.488	8	0.484	48	0.463
97	0.449	8	0.456	11	0.473	50	0.463
48	0.448	46	0.454	49	0.448	12	0.436
38	0.443	37	0.436	22	0.432	42	0.432
19	0.440	44	0.404	30	0.426	23	0.418
70	0.429	34	0.402	6	0.424	33	0.416
6	0.423	16	0.398	33	0.420	46	0.414
45	0.410	6	0.396	9	0.419	28	0.392
55	0.405	22	0.389	19	0.402	35	0.375
85	0.404	20	0.387	47	0.399	49	0.371
57	0.397	23	0.379	4	0.383	40	0.367
17	0.396	47	0.377	48	0.372	47	0.366
80	0.391	45	0.375	28	0.370	27	0.366
46	0.378	3	0.365	50	0.358	18	0.362
28	0.373	17	0.342	16	0.354	45	0.358
59	0.369	35	0.330	29	0.349	9	0.353
78	0.364	50	0.319	23	0.345	20	0.328
25	0.362	33	0.317	45	0.340	5	0.322
4	0.361	38	0.317	42	0.339	34	0.318
72	0.359	27	0.283	10	0.301	32	0.305
39	0.356	40	0.281	26	0.299	4	0.296
68	0.355	25	0.271	15	0.280	31	0.294
56	0.350	49	0.260	32	0.259	8	0.289
43	0.344	42	0.245	37	0.249	1	0.285
99	0.344	24	0.244	2	0.239	14	0.276
84	0.342	19	0.235	12	0.236	16	0.246
18	0.341	26	0.234	24	0.216	25	0.224
21	0.341	9	0.222	21	0.213	43	0.211
89	0.337	21	0.216	40	0.211	6	0.190
75	0.322	31	0.209	7	0.205	17	0.189
67	0.322	14	0.176	39	0.204	15	0.161
73	0.321	4	0.161	25	0.176	11	0.136
32	0.319	15	0.127	36	0.143	3	0.125
2	0.314	2	0.120	18	0.141	30	0.120
74	0.313	11	0.060	41	0.122	7	0.097
35	0.313	43	0.014	38	0.103	22	0.062
23	0.313	28	−0.009	20	0.089	2	0.055
33	0.310	30	−0.055	1	−0.070	10	0.023
81	0.310	29	−0.269	13	−0.119	21	−0.097

#	CL	#	CG	#	CV	#	CR
20	0.309						
65	0.308						
87	0.303						
15	0.301						
71	0.287						
8	0.286						
7	0.284						
3	0.282						
62	0.280						
90	0.267						
16	0.248						
36	0.243						
63	0.238						
41	0.233						
47	0.230						
100	0.221						
1	0.218						
96	0.209						
14	0.207						
29	0.205						
95	0.203						
13	0.202						
49	0.202						
79	0.201						
94	0.182						
12	0.177						
83	0.177						
37	0.168						
50	0.159						
22	0.156						
30	0.150						
31	0.147						
61	0.141						
34	0.136						
51	0.127						
24	0.116						
54	0.115						
26	0.113						
60	0.100						
11	0.083						
40	0.074						
53	0.057						
86	0.045						

#	CL	#	CG	#	CV	#	CR
5	−0.004						
91	−0.050						
10	−0.062						
89	−0.086						
76	−0.121						
27	−0.125						
66	−0.348						

Ⅲ. Test Method Facet (TMF) Analyses

Many studies have indicated that test results are influenced not merely by test-takers' ability but also by test methods (Bachman and Palmer, 1981, 1982; Shohamy, 1983, 1984; Bachman et al., 1989). Thus, for the sake of maximal validity, it is well worth the effort to investigate the extent to which the given test methods are reliable. The test method facets are analyzed through qualitative and quantitative approaches to verify the validity of test formats. The findings for each test are as follows:

1. Listening

The following are the reliability indices for each listening task pertaining to a subskill component.

(* under REL denotes relatively low reliability, hereafter.)

TASK	REL
(1) Picture Description	.541
(2) Identification	.315*
(3) Paraphrasing	.695
(4) One Sent.-Appr. Resp.	.686
(5) Dialogue-Appr. Resp.	.536
(6) Short Dialogue-Q & A	.695
(7) Discourse-Q & A	.510
(8) Long Dialogue-Q & A	.482*

1.1. Length

As mentioned earlier, the length of listening/reading passages does not increase reliability. They are in fact likely to impede test-takers' concentra-

tion, resulting in low discriminability and consequently low reliability. Even the native speakers confessed that the lengthy test (e.g., TOEFL Listening Test Section C) appeared to coerce their memory load heavily in addition to their language competence. Therefore, the last part in lengthy format with lowest reliability will be eliminated for the 2nd pilot test.

The 2nd part is to be discarded because of its low reliability. This is presumably due to the fact that the ability to reproduce the identical sentence does not have a significantly positive effect on the global listening ability. Thus, the number of items will be reduced to 80 items for increased practicability. This reduction in number is complemented by CG and CV based on the finding discussed in the Correlational Study.

1.2. Unadulterated Listening Task Only

The most important feature of SNUCREPT Listening Test is the 'Aural Mode Only' format. It is aimed at measuring aural comprehension only. As the choices as well as the listening passages and questions are presented aurally, the amount of aural input presented is much more than that of conventional listening tests. Compensating for the inherent problem of limited sampling in language testing, this method helps increase the reliability. 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 obvious that the test results vary significantly depending on what kind of test-taking strategy the test-takers choose to employ.

It is also worth noting that in the case of TOEFL or TOEIC while the directions for each part are being read for 30 seconds or so, the test-takers with 'test-wiseness' prepare for the test by reading the choices. However, as instructed by the directions, the naive test-takers read and listen to the instructions simultaneously. It is no more than common sense that this test-wiseness factor will significantly influence the performance of test-takers.

Hence, it is certain that this seemingly innocuous format does function as

a significant test bias factor. This case in point clearly demonstrates how significant an effect a test method factor can have on test-takers' performance.

1.3. Two Time Exposure to the Aural Passage

Like the time-honored 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. The question and answer type is most problematic in that its aritificiality 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 guess what is going on in a kidnapping incident!)

This test method is a controversial issue in that some argue that we do not listen to others twice in real life communication. In an adverse natural setting with a lot of noise interrruption, however, we do allow ourselves to ask our interlocuters to repeat what they have just said. This method also allows for limitation of listening test in space and time. That is, it is important to take into consideration the possibility of an adverse natural setting in which someone happens to cough or make noise loudly enough to deny the remaining test-takers a chance to listen properly. Thus, it is for the sake of fairness that this method should be seriously considered as it is with UCLES test batteries.

1.4. One-Passage-One-Item (OPOI) Principle

The OPOI principle is essential to local independence, one of the funda-

mental 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. Therefore, it is imperative that the OPOI principle be employed throughout the test.

As was mentioned earlier, lengthy listening passages as in TOEFL Listening Section C burden even native speakers with a heavy memory load. Even if we have to use our memory for successful aural comprehension, the language test should be designed to minimize the memory factor, one of the extra-linguistic factors. This also contributes to a serious test bias.

1.5. Diversity of Tasks

As has been documented many times, the test method does affect the test performance, leading to test method biases. It is only through employing a variety of valid tasks that we can preclude serious test method biases.

Some tasks inherently constitute valid methods to measure some subskills. For example, segmental-related subskill can be measured more appropriately by elemental tasks. Needless to say, pronunciation is best measured in aural mode. It is simply an invalid method to measure spoken language skill through a written form of test (최인철, 1991). Interestingly enough, however, there still exist anachronistic language tests which dare to claim their validity, in spite of the measurement of pronunciation with paper and pencil test format.

2. Grammar

The following are the reliability indices pertaining to each grammar test task.

TASK	REL
(1) Gap-filling: Sentence	.730
(2) Gap-filling: Dialogue	.640
(3) Error within Slashes	.665

(4) Choose Correct Sentence .323*
(5) Choose Wrong Sentence .336*

2.1. Choice of Un/Grammatical Sentence without Context: Formality

The lowest reliability indices are obtained from parts 4 and 5 which require the test-takers to select the grammatically correct or incorrect sentence. This can be explained by the native speakers who take the test. They describe the problematic aspect of this test method. Without due consideration of the context, the test is designed to focus on the grammaticality of four single sentences.

In many cases, however, the grammaticality of a sentence is closely intertwined with the context to which the sentence is attached. It should be noted that TOEFL Grammar Test seems to be based on the philosophy that a grammar test should be isolated from meaningful context, using heavily context-reduced vocabulary (novel words, proper nouns) (Savignon, 1986). This is, however, a serious violation of both authentic language use and the authenticity of language testing. Therefore, the native speakers have difficulty switching back and forth from the formality of each sentence, having to conceptualize the formality without context. Thus, this problematic test method should be removed for the 2nd pilot test development.

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

Part 3 has relatively low reliability. This part is a slightly improved format from the TOEFL Section 2 structure test format, in which only four parts (word or combination of words) are underlined, but the other parts are excluded as choices. This format is in line with the characteristic of discrete-point test, which is considered invalid for proficiency tests. It also 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).

The present test format is designed to divide a sentence into four parts with three slashes. It 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 is to provide a more context-embedded, i.e., slightly longer discourse or two-exchange dialogue. In a discourse type format, the clause should be considered as the basic chunk for parts separated with slashes. It should be borne in mind that this method can easily make the item difficult. In a dialogue type, one sentence is viewed as a basic unit.

2.3. Crucial Factor of Time Allocation

As was discussed in the rationale of SNUCREPT, the grammar test is designed to validly measure the test-taker's acquistion, not learning, through inactivating test-taking 'wiseness' strategy. It is to approximate a speed test by maximizing the speediness of the test (최인철, 1991).

The informal survey of Korean test-takers indicates that the 25-minute-long TOEFL Grammar Test is far from a speed test for the majority of Korean students, at least. Therefore the validity of the grammar test results of TOEFL (which is supposed to measure genuine grammatical competence that can be used for communicative purpose) can be called into question in further research.

2.4. Gap-Filling Task

Gap-filling tasks are found to be fairly reliable regardless of the content domain, whether spoken English or written English. Based on aforementioned findings, this format will be used in addition to the error-detection task for the SNUCREPT.

3. Vocabulary

The following are the reliability indices pertaining to each task of the vocabulary test.

TASK	REL
(1) Paraphrasing	.478*
(2) Synonym	.392*
(3) Gap-filling: Dialogue	.647
(4) Gap-filling: Sentence	.763

3.1. Paraphrasing Task

The paraphrasing task has relatively low reliability. This task will be eliminated and replaced by gap-filling tasks, which are shown to be more reliable than the first two tasks. Furthermore, gap-filling tasks force the test-takers to refer to the context to answer a problem, which is essential to authentic language use and effective language acquisition.

3.2. Synonym of Underlined Part

This task has the lowest reliability index, which is quite understandable considering the process which the test-takers go through in taking this test method. As the test-takers are not required 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 Section 3 Vocabulary test uses this typical test method, which is of the discrete-point test format in that it requires the test-takers to focus on meticulous points regardless of context. This problematic method should be replaced by gap-filling tasks.

3.3. TOEFL Vocabulary Test

In addition to the problem with the TOEFL Vocabulary test format, the test also resorts to heavy use of context-reduced words (e.g., proper nouns, novel words) as in TOEFL Grammar test (Savignon, 1986). This is far from desirable in language testing as well as language teaching.

3.4. Gap-Filling Task

The gap-filling task proves to be fairly reliable regardless of the content domain, whether spoken English or written English. Based on aforementioned results and this finding, this format will be used for the SNUCREPT.

4. Reading

The following are the reliability indices for each reading subskill component and task pertaining to it.

TASK	REL
(1) Detail Q & A	.414*
(2) Main Idea/Topic Q & A	.571
(3) Inference Q & A	.811
(4) Coherence	.676
(5) Gap-Filling	.789
(6) Q & A	.730

4.1. 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-one of the essential assumptions of IRT. This principle helps us 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). 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 be designed to minimize the influence of this factor on the test performance.

4.2. Reasonably Diverse Task Formats

To preclude the task format bias, the SNUCREPT Reading test employs diverse test methods which are shown to be valid for measuring specific content domains.

4.3. Gap-Filling Task vs. Question-and-Answer Task

The gap-filling task proves to be very reliable. It is probably due 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 artificially developed by the test writers. Unlike other multiple choice formats, however, this format does not require the test writer to de

velop his own key since it is already provided in the original text. The fact that original text is used for developing keys maximizes the test authenticity and minimizes the artificiality of developing the text of keys. It is this characteristic that maximizes the reliability of a test (최인철, 1991).

Since the gap-filling method does not require the reader to focus on specific information, this method is well suited to measuring the ability to grasp the gist of the passage or to resort to inference. As the question-and-answer task can easily direct the readers to focus on specific information, this method is employed to measure the ability to identify detail information.

4.4. Coherence-related Task

The coherence-related tasks including deletion, insertion, and reordering are found to be fairly reliable. Native speakers' insight reveals, however, that deletion is the best method for measuring the competence regarding coherence. The other two methods tend to induce ambiguous keys, depending on the viewpoint and the logic of a reader (even with native reading ability), whereas the deletion method is likely to produce an item with a clear-cut key.

4.5. Inference Task

The inference task has fairly high reliability, even though the facility indices are generally low for the items. The inherent difficulty of the items should be take into account in distributing the items in a test.

4.6. Detail Information Task

The task measuring the ability to identify detailed information was the lowest, which can be attributed to the fact that this task tends to manifest itself as being a discrete-point test. Despite the lowest reliability, this method should be included in the SNUCREPT because the ability to identify detailed information is considered one of the essential elements of reading skill.

4.7. Main Idea/Topic Task

The question-and-answer task used to measure the ability to understand

the main idea/topic proves to be reliable. Some of the items measuring the ability to grasp the main idea and topic employ the gap-filling method.

Ⅳ. Item Response Theory (IRT)

1. Assumptions for Appropriate Application of IRT

Assumptions of IRT should be checked to ensure valid application to the SNUCREPT data. The results of checking unidimensionality are provided in the next section. The assumption of local independence is ensured through one-passage-one-item principle. Considering the sample size and test length, BILOG is employed and 2 parameter logistic (PL) model is chosen among the popular IRT models (Choi, 1991).

2. Unidimensionality Check

Stout's Factor Analysis-based unidimensionality check shows that all the tests are 'essentially' unidimensional at alpha=.01 (Stout et al., 1991). This finding thus validates the IRT modeling for the present study.

The P-values (probability index) show that the Grammar (.955633) and the Vocabulary (.914352) tests prove to be almost unidimensional. On the other hand, the Reading test has the P-value of .051869, which shows the degree of complexity of cognitive process required for taking this test.

The fairly high P-value of .496277 of the Listening test suggests that the test does not require cognitive processes as complex as the Reading test. This may be due to the fact that test-takers have more difficulty (as shown in descriptive statistics and IRT statistics) in taking the Listening test (which was believed to require far more complex cognitive process), thus utilizing their cognitive process as fully as in taking the Reading test.

Stout's Factor Analysis-based Unidimensionality Check

Test	T	P-value	Dimensionality
Listening	.009333	.496277	Essentially Unidimensional
Grammar	−1.702117	.955633	Essentially Unidimensional
Vocabulary	−1.368050	.914352	Essentially Unidimensional
Reading	1.626995	.051869	Essentially Unidimensional

V. IRT-based Test and Sample Statistics

1. 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. As the following table demonstrates, the IRT discrimination and difficulty indices have more precise and widely spread-out distribution than do the classical ones.

a and b indices estimated by 2 PL model are presented as follows:

SNUCREPT Item IRT Indices (a, b) sorted in descending order

	CL		CL		CG		CG		CV		CV		CR		CR
#	a	#	b	#	a	#	b	#	a	#	b	#	a	#	b
68	0.947	5	-3.120	39	1.255	14	-4.572	31	1.122	4	-2.277	26	1.147	1	-3.485
42	0.908	3	-2.950	10	1.025	2	-3.807	34	1.026	3	-2.218	39	0.902	13	-2.889
23	0.756	7	-2.493	32	0.924	6	-2.715	3	0.902	6	-1.500	19	0.759	12	-2.600
46	0.753	2	-2.033	36	0.863	9	-2.657	5	0.886	31	-1.499	24	0.717	16	-2.387
18	0.737	6	-1.650	41	0.775	22	-2.620	4	0.797	21	-1.234	41	0.618	4	-2.202
88	0.713	1	1.356	7	0.718	4	-2.488	35	0.766	15	-1.176	28	0.607	6	-2.048
7	0.696	13	-1.352	13	0.671	7	-2.121	14	0.622	2	-1.165	33	0.603	9	-1.958
25	0.679	4	-1.204	5	0.628	5	-2.036	6	0.617	7	-1.128	25	0.600	29	-1.868
64	0.672	42	-1.040	18	0.606	10	-1.871	11	0.602	34	-1.106	37	0.600	26	-1.724
4	0.665	31	-0.999	37	0.573	31	-1.846	44	0.569	17	-0.971	36	0.598	28	-1.601
9	0.636	40	-0.991	1	0.570	1	-1.658	9	0.561	5	-0.891	38	0.568	5	-1.508
3	0.585	85	-0.896	8	0.525	8	-1.611	43	0.567	14	-0.805	16	0.514	19	-1.476
82	0.574	68	-0.747	46	0.522	23	-1.436	46	0.539	19	-0.800	13	0.500	15	-1.423
92	0.573	28	-0.704	48	0.498	3	-1.431	19	0.530	27	-0.698	44	0.486	25	-1.256
77	0.566	9	-0.560	6	0.494	18	-1.389	27	0.464	25	-0.667	48	0.478	17	-1.237
58	0.555	77	-0.461	12	0.481	15	-1.248	17	0.457	35	-0.658	35	0.464	36	-1.218
43	0.546	19	-0.426	44	0.467	13	-1.225	30	0.445	20	-0.628	20	0.446	24	-1.007
40	0.535	56	-0.404	34	0.458	24	-1.112	29	0.436	11	-0.410	50	0.446	37	-0.940
72	0.531	33	-0.376	47	0.452	16	-1.016	16	0.414	9	-0.295	23	0.445	41	-0.899
6	0.531	23	-0.372	45	0.436	12	-0.981	47	0.412	28	-0.288	12	0.422	33	-0.791
75	0.529	25	-0.338	20	0.433	32	-0.936	28	0.393	32	0.096	42	0.414	27	-0.787
13	0.528	67	-0.319	23	0.427	36	-0.350	26	0.388	24	0.263	45	0.406	18	-0.681

#	CL a	#	CL b	#	CG a	#	CG b	#	CV a	#	CV b	#	CR a	#	CR b
93	0.507	65	-0.300	16	0.415	39	-0.067	22	0.374	43	0.326	9	0.406	14	-0.551
85	0.504	88	-0.279	22	0.405	19	0.110	33	0.363	29	0.357	29	0.398	23	-0.521
22	0.504	10	-0.268	38	0.398	37	0.178	2	0.342	26	0.452	46	0.395	31	-0.518
20	0.498	18	-0.224	17	0.393	27	0.190	23	0.339	44	0.551	18	0.392	8	-0.355
74	0.495	20	-0.163	50	0.377	20	0.207	15	0.338	12	0.553	5	0.382	39	-0.275
61	0.494	55	-0.157	3	0.361	34	0.498	21	0.336	30	0.568	8	0.381	20	-0.271
55	0.487	95	-0.132	33	0.355	41	0.499	42	0.314	16	0.619	40	0.373	44	-0.046
16	0.487	64	-0.107	35	0.335	35	0.649	12	0.294	46	0.621	34	0.369	38	0.027
21	0.482	46	-0.098	40	0.322	33	0.850	10	0.276	22	1.463	49	0.356	46	0.172
19	0.479	48	0.001	26	0.312	44	0.955	48	0.273	10	1.481	14	0.352	35	0.287
1	0.465	43	0.054	25	0.311	46	0.995	49	0.268	42	1.491	47	0.342	11	0.336
59	0.465	24	0.073	49	0.303	25	1.100	7	0.268	33	1.576	4	0.334	40	0.381
48	0.461	16	0.122	42	0.295	17	1.172	50	0.266	23	1.682	27	0.321	49	0.426
80	0.461	47	0.220	19	0.271	48	1.224	24	0.248	49	2.021	1	0.316	30	0.467
5	0.455	72	0.240	27	0.270	21	1.374	25	0.241	37	2.156	32	0.316	7	0.527
67	0.454	57	0.256	9	0.269	45	1.385	37	0.233	40	2.559	17	0.303	34	0.659
57	0.446	93	0.311	24	0.261	38	1.561	45	0.219	48	2.606	22	0.287	45	0.672
38	0.444	75	0.349	31	0.257	47	1.882	41	0.215	47	2.749	15	0.274	32	0.684
33	0.442	22	0.351	21	0.251	26	2.050	40	0.205	41	3.505	7	0.258	50	1.092
2	0.440	61	0.398	4	0.225	11	2.256	20	0.204	50	3.553	31	0.256	42	1.507
39	0.438	30	0.403	29	0.208	42	2.418	38	0.190	1	4.251	6	0.226	48	1.590
28	0.428	15	0.410	14	0.205	40	2.506	39	0.185	36	4.294	11	0.218	22	1.780
52	0.421	38	0.436	15	0.198	50	2.676	32	0.183	38	4.308	2	0.194	2	1.944
95	0.412	14	0.467	2	0.180	49	2.853	36	0.168	39	5.483	3	0.186	47	2.032
44	0.395	96	0.488	43	0.173	28	4.763	18	0.155	45	5.766	43	0.178	3	2.058
56	0.394	59	0.494	30	0.150	43	5.457	1	0.153	18	6.805	21	0.118	43	4.938
65	0.394	39	0.549	11	0.148	30	8.891	13	0.131	13	12.331	10	0.104	10	7.941
47	0.375	21	0.692	28	0.137	29	10.982	8	0.102	8	15.862	30	-0.138	21	10.545
31	0.365	82	0.698												
99	0.365	11	0.748												
87	0.355	92	0.759												
24	0.353	74	0.835												
35	0.345	29	0.851												
53	0.336	73	0.852												
30	0.331	52	0.935												
96	0.330	53	0.999												
78	0.318	34	1.029												
71	0.318	80	1.070												
70	0.313	8	1.131												
15	0.303	35	1.227												
14	0.291	99	1.381												
73	0.281	44	1.401												
29	0.276	36	1.440												
62	0.272	58	1.454												
100	0.272	78	1.547												
17	0.252	17	1.591												

CL #	CL a	CL #	CL b	CG #	CG a	CG #	CG b	CV #	CV a	CV #	CV b	CR #	CR a	CR #	CR b
8	0.248	87	1.741												
45	0.239	70	1.778												
84	0.238	97	1.840												
36	0.231	86	1.867												
81	0.223	83	1.963												
54	0.223	100	1.987												
90	0.214	60	2.202												
97	0.210	45	2.251												
12	0.210	37	2.264												
10	0.203	50	2.438												
63	0.201	71	2.451												
60	0.198	41	2.564												
26	0.197	27	2.837												
50	0.194	62	2.845												
34	0.192	91	2.883												
69	0.189	84	3.123												
51	0.172	51	3.193												
32	0.172	90	3.203												
11	0.170	26	3.245												
83	0.170	12	3.518												
89	0.162	81	3.564												
41	0.159	54	4.323												
37	0.157	49	4.496												
94	0.150	89	4.925												
86	0.144	32	5.012												
91	0.141	63	5.335												
27	0.134	79	6.221												
79	0.129	76	6.518												
76	0.126	94	6.537												
49	0.116	66	6.718												
66	0.087	69	6.941												
98	0.075	98	9.635												

2. Sample Statistics

2.1. More Precise Measurement

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 show

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 (as underlined in the following table) within IRT framework. The ability level estimates range from −4 to +4 in the IRT application software BILOG output. 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.

SNUCREPT Reading Test Score

ID#	Content Component	# Tried	# Right	CTT Percent	IRT Ability
88630743					
	DETAIL	10	7	.7000	.7250
	MAIN	10	9	.9000	2.1505
	INFER	25	17	.6800	.6830
	COHER	5	2	.4000	.4077
	TOTAL	50	35	.7000	.8974
89660713					
	DETAIL	10	8	.8000	1.6811
	MAIN	10	10	1.0000	4.0000
	INFER	25	16	.6400	−.0348
	COHER	5	0	.0000	−1.2580
	TOTAL	50	34	.6800	.3886
92670693					
	DETAIL	10	6	.6000	1.0614
	MAIN	10	8	.8000	.8580
	INFER	25	19	.7600	1.2543
	COHER	5	2	.4000	.5454
	TOTAL	50	35	.7000	1.1340

2.2. Diagnostic Purpose

IRT also allows us to estimate the ability level of each test-taker on each subtest. This feature serves the diagnostic purpose as well as the certification in a more specific subskill component or content domain. A careful analysis of the results can provide valuable information for language education programs as well as for individual language learners. Refer to the tables given above.

2.3. Grade/ Certification

Each test carries different weights depending on their saliency in terms of language use and functions. TOEFL, however, provides a lump score for Section 3 which combines vocabulary and reading, which means it gives the same weight to vocabulary and to reading. Because completing a vocabulary item consumes much less time than a reading item which involves vocabulary ability, it is against common sense in testing to give them equal weight. Therefore, SNUCREPT employs a proportional weighting system to give different weights in calculating scores.

This weighting system provides the transformed score in a report form ranging from 100 minimum points up to 900 maximum points. 9-class-grades (A+, A, B+, B, C+, C, D+, D, F) or certification (pass/failure) are to be given on the basis of large-scale norm data within this scoring system.

3. Computer Adaptive Test (CAT)

An answer to the most critical issue of reliabilty and validity can be solved by the introduction of the computer adaptive test, or CAT for short (Choi, 1989). Since the concept of reliabilty and validity is a parallel function of item difficulty and test-takers' ability, it is imperative that item difficulty is matched to the ability level. This can be easily manipulated by CAT which has many advantages in terms of security, economy, precise measurement, and overall reaction to the test format on the part of test-takers. Our ultimate goal is to apply CAT to the real world language testing environment.

VI. Correlational Analysis on TSE & SNUCREPT

Correlational analyses on Test of Spoken English (TSE) and SNUCREPT are intended to validate SNUCREPT Vocabulary and Grammar Speed Test. The table is as follows.

Correlation Bet. TSE & CREPT

	TSE (Test of Spoken English)				CREPT			
Test	Pronun	Grammar	Fluency	Comprehen	CL	CG	CV	CR
TG	.2784	—						
TF	.5292**	.4064**	—					
TC	.3334*	.6035**	.5983**	—				
CL	.5465**	.3859*	.3441*	.4631**	—			
CG	.1803	.6759**	.4930**	.6326**	.6142**	—		
CV	.1216	.3725*	.4288*	.5198**	.6659**	.4707**	—	
CR	.1342	.2765	.3411	.3496	.4450*	.3839*	.4855*	—

(N of cases: 86 1-tailed Signif: * — .01 * * — .001)

1. High Correlation between TSE and CG & CV

The results show a high correlation between TSE (especially fluency and overall comprehensibility) and CG and CV tests. This finding strongly supports the rationale of language tests being speed tests.

2. High Correlation between CL & CG & CV

A high correlation between CL, CG and CV also suggests that the speed grammar and vocabulary tests can complement the listening test. This finding can justify the reduction of 100 listening items to 80 in terms of practicality (time length) and construct validity.

Ⅶ. Conclusion

The overall results demonstrate that the systematic test development is essential to fulfilling the adequacy of validity and reliability required for desirable tests. The whole process for construct validation involved qualitative and quantitative approaches pertaining to each issue in question, e.g., 1) systematic test development (including content analysis), 2) analysis of descriptive statistics and test/item indices (including reliability, facility, and discriminability), 3) item/distractor analysis, 4) quantitative and qualitative approaches to test method analysis, 5) IRT application for more precise measurement (including Factor Analysis-based dimensionality check), and 6) correlational analyses.

The above findings lay a fundamental basis on which to revise the format and the number of items for the second pilot test, which is to be further validated to finalize the testing procedures and specifications of SNUCREPT. The table of specification for the 2nd pilot test is given as follows.

Table of Specification for 2nd Pilot Test

	# Items	Length (min)	Weight	Total (pts)
Listening:	80	40	4	200
Grammar:	40	10	1	50
Vocabulary:	40	10	1	50
Reading:	40	40	4	200
Total:	200	100		500 pts

The SNUCREPT-TOEFL comparability study will be systematically done with the 2nd pilot test development to investigate the extent to which SNUCREPT and TOEFL can be comparable in terms of test content domain and format. IRT equating will be employed, where appropriate, to develop a transformation formula for a practical purpose to predict TOEFL score from SNUCREPT score, or SNUCREPT from TOEFL. The 2nd pilot test currently being developed will provide more conclusive evidence on which to base the investigation of the comparability of the two tests in terms of CLA and TMF. It is hoped that the present study sheds some light on how to apply the empirical evidence and the theoretically sound frameworks to the overall processes of test development and construct validation in language testing.

References

최인철(1991) '언어테스팅의 이론과 실제,' 정동빈 편, 영어 교육론 (265-299), 서울: 한신문화사.

Bachman, Lyle F. (1986) 'The Test of English as a Foreign Language as a Measure of Communicative Competence,' In Charles W. Stansfield (ed.), *Toward Communicative Competence Testing: Proceedings of the Second TOEFL Invitational Conference*, May, 1986, ETS.

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

Bachman, Lyle F. and Palmer Adrian (1981) 'The Construct Validation of the FSI Oral Interview,' *Language Learning* 31.1, 67–86.

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 (in press) *An Investigation into the Comparability of Two Tests of English as a Foreign Language: The Cambridge-TOEFL Comparability Study*, UCLES.

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 (1988) 'The Necessity of Teaching English Fast Speech Phenomena for Better Aural Comprehension Skills in the Korean Context,' Master's Thesis, Urbana, IL: University of Illinois at Urbana-Champaign.

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 Second Language Acquisition: Application of Item Response Theory to Language Testing*, New York: Peter Lang Publishing Inc.

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

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.

Dubin, F, Eskey D. E., and Grabe W. (1986) *Teaching Second Language Reading for Academic Purposes*, Addison-Wesley Publishing.

Goodman, Kenneth S. (1967) 'Reading: A Psycholinguistic Guessing Game,' *Journal of the Reading Specialist* 4, 126–135.

Gronlund, Norman (1985) *Measurement and Evaluation in Teaching*, (5th ed.) NY: Macmillan Publishing Company.

Krashen, Stephen D. (1985) *The Input Hypothesis*, London: Longman Inc.

Lado, Robert (1961) *Language Testing*, NY: McGraw-Hill.

Madsen, Harold S. (1983) *Techniques in Testing*, NY: Oxford Univ. Press.

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

Richards, J. C. (1985) *The Context of Language Teaching*, NY: Cambridge Univ. Press.

Rivers, Wilga and Mary Temperley (1978) *A Practical Guide to the Teaching of English*, NY: Oxford Univ. Press.

Savignon, Sandra J. (1982) 'Dictation as a Measure of Communicative Competence in French as a Second Language,' *Language Learning* 32, 33–51.

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*, May, 1986, ETS.

Shohamy, Elana (1983) 'The Stability of Pral Proficiency Assessment on the Oral Interview Testing Procedures,' *Language Learning* 33, 527–40.

Shohamy, Elana (1984) 'Does the Testing Method Make a Difference? The Case of Reading Comprehension,' *Language Testing* 1.2, 147–70.

Smith, Frank (1982) *Understanding Reading: A Psycholinguistic Analysis of Reading and Learning to Read*, (2nd ed.) NY: Holdt, Rinehart and Winston.

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

Ur, Penny (1984) *Teaching Listening Comprehension*, Cambridge: Cambridge Univ. Press.

Language Research Institute
Seoul National University
Shillim-dong, Kwanak-ku
Seoul 151-742
Korea