Review of Seoul National University
Criterion-Referenced English Proficiency Test
(SNUCREPT)

Lyle F. Bachman

1. Introduction

The Seoul National University Criterion-Referenced English Proficiency test (SNUCREPT) was developed "to measure of overall English proficien­cy of SNU students and the larger public" (Choi 1993). I had access to the following materials for this review:

- Examinee materials for SNUCREPT 4
- Tape script for Listening Comprehension part of SNUCREPT 4
- "A Guideline for Writing Test Items" (for Reading Comprehension and Listening Comprehension parts)
- Several published reports about the SNUCREPT (see References, below)

2. Description of the Test

SNUCREPT 4 is a 200 items English placement aimed at assessing two areas of language use (listening comprehension and reading comprehen­sion) and two areas of language knowledge (grammar and vocabulary). The listening comprehension section of the test, which is administered au­rally via tape, includes 60 multiple-choice items divided into five parts. Part I includes 10 items, each consisting of a single-sentence stem and four single sentence options, one of which is "closest in meaning" to the sentence in the stem. Part II includes 10 items in which the stem includes the first part of a one-turn conversational exchange, with the appropriate second part of the exchange given in one of the four choices.
includes 10 items in which the stem includes the first turn and the first part of the second turn of a two-turn conversational exchange with the second half of the second turn given in one of the four choices. Part IV includes 15 items that repeat this pattern, but with three- or four-turn conversations. In Part V, there are 15 items, each consisting of a short piece of non-interactive oral discourse, followed by a question, with the "best" answer to the question provided in one of the four choices.

The reading comprehension section of SNUCREPT 4 consists of 35 multiple-choice items differentiated into parts on the basis of their different task types. The first part consists of 15 short written texts, each followed by a single question, the "best" answer to which is given in one of the four choices. The second part consists of 15 short "gap-filling" texts in which a word or phrase has been deleted, with the word or phrase that best fills the gap provided in one of the four choices. The last part consists of five close-elide texts, in which one of four phrases must be deleted to provide "the best coherent meaning."

The grammar and vocabulary sections of SNUCREPT 4 include fairly conventional multiple-choice completion-type items in which the "correct" choice (grammar) or "best" (vocabulary) choice is provided among the four options. There are 50 such items in each of these two parts.

2.1. Design Features

The SNUCREPT incorporates both applied linguistic and psychometric principles in its design. On the applied linguistic side, the constructs to be measured are derived from current models of communicative language ability (e.g., Bachman, 1990; Canale, 1983; Cummins, 1983), while the features the test tasks, or items, incorporate facets from Bachman's (1990) test method facets (TMF) framework. These features are specified explicitly in "A Guideline for Writing Test Items", for the listening comprehension and reading comprehension subtests. For listening comprehension, these guidelines clearly specify the functions (e.g., requesting and offering help, expressing opinion), notions or topics (e.g., invitations, appointments, natural science, arts), and situations (e.g., international travel, telephone conversation, news report, advertisement) that are included in the content of the test tasks, the mode of language use (interactive vs. non-interactive),
and the language components (e.g., pronunciation-linking, BICS expression (Cummins, 1983), detailed information, main idea, and inference) to be measured. The listening section also incorporated the feature that all input is presented aurally only, thus eliminating the possibility that listening test results will reflect test takers’ reading ability (Choi 1994, p. 328). In order to maximize the interactiveness (Bachman & Palmer 1996) of the listening tasks, test takers listen to the listening texts once in a “macro-listening manner” and then to the question, in order to understand what they are expected to listen for, and then listen to the text a second time in a “micro-listening manner” in order to listen for specific information (Choi 1994, p. 329).

The reading comprehension section is equally well-specified, in terms of topic (non-academic/practical and academic/non-esoteric), task type (question-answer, gapfilling, deletion), and language subcomponents to be measured (detailed information, main idea/topic, inference, cohesion and coherence). All parts of the test impose time limitations, which are intended to reflect the “speediness” that is characteristic of “real-time communicative settings,” so as to be able to make inferences about acquired, as opposed to learned, communicative competence (Choi 1994, p. 312).

On the psychometric side, SNUCREPT is developed according to principles of criterion-referenced (CR) measurement and item-response theory (IRT). As has been pointed out by a number of measurement specialists (e.g., Nitko 1984) there are several approaches to specifying the criterion for a given CR test, one of which is to specify the domain of the content or ability that is to be measured, in which case we can refer to the test as a “domain-referenced” test. If we can consider the specifications above as specifying a domain, then SNUCREPT can be considered to be a domain-referenced CR test, with the construct domain specified by the language components and the domain of generalization specified by the TMFs, or task characteristics.

In order to help assure that the IRT requirement of local independence is satisfied, both the listening and reading comprehension parts of the test incorporate what Choi (1993, 1994) calls the “one-passage-one-item principle.” Local independence implies that a given test taker’s performance on a given test task is independent of how she or he performs on other test tasks. Choi (1994) points out that when two or more items are based on
the same listening or reading passage, these are likely to be interdependent and are thus detrimental to test fairness and hence, to validity. This is of particular concern with passages that deal with technical topics. The relatively short texts in the SNUCREPT listening subtest are also defined to lessen the potential effect of memory on test performance.

Item difficulty is another design feature, with items specified to fall into five levels of difficulty (Choi 1993). According to “A Guideline for Writing Test Items”, item writers are expected to construct items according to the content specifications described above, and with a specified range of difficulty levels, approximating, a normal distribution. According to the results of the first pilot test, reported by Choi (1993), both content and statistical criteria are used in the selection of items, so that the majority of item facility, values fall within the range of $0.2 < p < 0.8$. (listening: 84%; grammar: 78%; vocabulary: 76%; reading: 78%), while items that are judged to measure important areas of content are included even though they might have low or even negative discrimination indices. From this it is clear that both content and statistical considerations are taken into account in the construction and selection of items.

2.2. Score Distributions

Choi (1993) reports the descriptive statistics for the scores on the first pilot SNUCREPT, and these indicate that scores are distributed reasonably normally, with means at around 50%. The standard deviations indicate reasonably wide dispersions, while the skewness and kurtosis indicate that the score distributions are essentially symmetrical and mesokurtic, as would be characteristic of a normal distribution.

3. Measurement Qualities

Bachman & Palmer (1996) define the overall usefulness of a language test in terms of six qualities: reliability, construct validity, authenticity, interactiveness, impact and practicality—and suggest that this definition of usefulness provides a framework for evaluating the extent to which a given language test is useful for its intended purposes. On the basis of the materials to which I had access, I can address the qualities of reliability and con-
struct validity in terms of evidence provided, and offer some opinions about the test's potential for authenticity and interactiveness.

3.1. Reliability

Choi (1994) reports norm-referenced (NR) reliability estimates—alpha coefficients—for all four subtests that are above .80 (listening: .857, grammar: .854, vocabulary: .825, reading: .861). Standard errors of measurement based on the alpha coefficients are also reported, as is recommended by the APA Standards (American Psychological Association 1985). While these reliability estimates are at acceptable levels, given the intended use of the test, it is well-known that coefficient alpha may overestimate reliability in the cases where tests are overly speeded, and it thus would be advisable for the test developers to utilize additional procedures (e.g. test-retest or parallel forms) that are less sensitive to the potential effects of speededness, for estimating reliability. Furthermore, since SNUCREPT is designed to be a CR test, it is important that the test developer also provide CR analogues of NR reliability estimates, that is, estimates of dependability and decision consistency at cut scores, as described in the CR literature, (e.g. Berk, 1984; Brennan, 1984).

3.2. Construct Validity

According to Messick, validity "is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support... inferences and actions based on test score." (1989, p. 13). Test validation thus involves bringing together a variety of types evidence and developing a cogent argument for the relevance of this evidence in supporting the interpretations and uses for which the test is intended. Choi (1993, 1994) provides a wealth of such evidence, clearly describing each and discussing its relevance to the validity of the interpretations of SNUCREPT scores.

Evidence for content relevance and coverage. Evidence provided in Choi (1994) for content relevance and coverage in support of the validity of score interpretations consist of a detailed discussion of the content considerations incorporated into the SNUCREPT design, which have been discussed above, as well as a test method facets (TMF) analysis based on Bachman's (1990) model. This analysis describes in detail the characteristics of rubric
(instructions, time allocation, test structure) and input (length, channel, format, language, propositional, or topical content, degree of contextualization).

An important aspect of the argument for content relevance is investigating the relationship between the content of test items and their measurement properties. One approach to this is to use indicators of the TMFs of test items as indicators of test content, looking at the correlations among these, as well as the correlations between these indicators and the difficulty and discrimination of these items. Drawing on the work of the Cambridge-TOEFL Comparability Study (Bachman, Davidson, Ryan and Choi, 1995), Choi (1994) develops indicators of a number of item TMFs, such as total number of words and sentences, as indicators of length, number of clauses per sentence, as an indicator of grammatical complexity, ratio of abstract to content words, as an indicator of degree of abstractness, and ratings of academic/nonacademic/culture specific topics as an indicator of topical content, and uses these as variables in a correlational analysis. Also included in the analysis are the item writer’s predicted difficulty of the item and the Flesh and Fog readability indices. The correlations between predicted difficulty and the 2-parameter IRT and classical difficulty indices are quite high, indicating a high degree of accuracy in item writers’ prediction of item difficulty. High correlations were also observed between the readability indices and other indicators of TMFs, while correlation between the readability indices and item difficulty were very low, suggesting that the readability indices do not contribute much information that is not already captured in the other TMF variables. Finally, significantly high correlations of indicators of grammatical complexity, degree of abstraction and topical content with item difficulty suggest that these TMFs may be particularly important in the design of the test.

While these findings provide supportive evidence for content relevance, a more powerful statistical procedure, multiple regression analysis, should, in my view, also be used to investigate the relationships between the TMFs as predictors and item difficulty and discrimination as criteria, as has been illustrated in a study by Bachman, Davidson and Milanovic (1996). I would also suggest that these relationships be investigated separately for each of the four subtests.

Evidence for convergence and discrimination of scores. Another type of evi-
vidence for the construct validity of interpretations is evidence demonstrating that scores on a test are positively related to other measures of the same construct (convergence) and not positively related to measures of different constructs (discrimination). Choi (1994) investigates this in two ways: an IRT dimensionality check and correlational analysis. A standard approach to IRT dimensionality check (Stout, Nandakumar, Junker and Chang, 1991) indicates that each of the four subtests of the SNUCREPT is essentially unidimensional, which suggests that the items in each subtest are measuring essentially the same construct. This finding has been replicated in a subsequent study conducted with another group of subjects and with another form, the 4th Pilot form, of the SNUCREPT (Choi 1995). In the other approach used, correlational analysis, it was found that the correlations among the SNUCREPT subtests and those of the Test of Spoken English (TSE) were examined to investigate both convergent and discriminant evidence of construct validity. Significantly high correlations between the SNUCREPT and TSE scores of grammar (.676) and listening/comprehensibility (.463), as well as between the SNUCREPT listening and TSE pronunciation scores (.547) provide evidence of convergence (Choi 1994). Similar patterns of convergence were observed in a study aimed at comparing the SNUCREPT and the Test of English for International Communication (TOEIC) (Choi 1995).

While there was considerable evidence for convergence among these correlations, the high correlations of the SNUCREPT grammar and vocabulary scores with the TSE comprehensibility scores (.633 and .520, respectively) and TSE fluency scores (.493 and .429, respectively) suggest a lack of discrimination among these constructs. The fact that the highest correlations among the SNUCREPT subtests are those between the listening comprehension scores and the grammar and vocabulary scores (.614 and .666, respectively) is also suggestive of lack of discrimination. Similar evidence of non-discrimination between grammar and vocabulary scores and scores on measures of other constructs was also evident in the correlations between SNUCREPT and TOEIC scores (Choi 1995). These high correlations between the SNUCREPT grammar and vocabulary subtests and measures of other components of language ability indicates a lack of discrimination these test scores, which the test developers might well consider investigating further, with an eye to possibly dropping some subtests entirely (i.e.,
grammar, vocabulary) and revising others (i.e., listening comprehension) significantly so as to provide more distinct measures of the constructs they are intended to measure.

3.3. Authenticity

Bachman & Palmer (1996) define authenticity as the degree of perceived correspondence between characteristics of the test tasks and situation and features of non-test language use tasks and situations. Further, they point out that authenticity should be considered to be a relative quality, so we can only characterize a given test task as either more or less authentic than another test task, and not as “authentic” or “inauthentic”. Several features of the SNUCREPT contribute to the relative authenticity of various types of tasks. For example, presenting both listening input and questions aurally in the sections of the listening comprehension subtest that involve conversational exchanges, clearly corresponds more closely to non-test language use than if the questions were presented in writing, to be read. Similarly, the use of original texts in the reading part of the test adds to the authenticity of this subtest. Finally, the fact that all parts of the test are designed to reflect the “speediness” that is characteristic of “real-time communicative settings” contributes to the authenticity of the test.

At the same time, there are features of the test that tend to lessen its authenticity. The parts of the listening comprehension subtest that require test takers to select an appropriate paraphrase or to guess the setting in which a conversation is likely to have taken place, for example, represent tasks that test takers are seldom, if ever, likely to encounter outside of the test itself. The least authentic parts of the SNUCREPT, in my view, are the grammar and vocabulary subtests, since these focus on formal aspects of the language itself, rather than on communicative value, or meaning, as is generally the case on non-test language use. Moreover, the tasks in these subtests are themselves decontextualized, and taken as a set, do not constitute discourse, as this is generally understood by applied linguists. For this reason, as well as that mentioned above, with respect to lack of discriminant validity, I would recommend that the developers seriously consider dropping these subtests. In addition, the authenticity of the test overall might be improved if the test developers moved away from multiple-choice questions and included item types, such as short-answer completion tasks,
that require test takers to produce language in their response. Ultimately, of course, authenticity is a matter of perceptions, and the test developers thus might want to consider collecting feedback from test takers and test users about the perceived authenticity of the test.

3.4. Interactiveness

Bachman & Palmer (1996) define interactiveness as “the extent and type of involvement of the test taker’s individual characteristics in accomplishing the test task (p. 25)”. In a language test, the individual characteristic that we are most interested in is language ability. With respect to the interactiveness of SNUCREPT tasks, in the absence of observational evidence that pertains to interactiveness, one can only speculate on their potential interactiveness. The test developers are clearly aware of the importance of interactiveness, however, and mention several features of SNUCREPT that pertain to this test quality. Presenting the aural input to test takers twice in the question and answer section of the listening comprehension subtest, for example, is intended to provide test takers greater opportunity to interact with both passages and the questions that follow. In addition, the developers have made revisions to the test since its inception that are aimed at increasing the test’s potential for interactiveness. Two item types, one from the grammar subtest (selecting the grammatically correct sentence in isolation), and one from the vocabulary subtest (selecting the closest synonym for an underlined word in the stem), which were included in the first pilot version of SNUCREPT, have since been dropped, in part, because of their low potential for interactiveness.

4. Summary

The SNUCREPT is a well-designed and carefully researched EFL proficiency test. It is soundly grounded in both applied linguistic and psychometric theory. The content and task characteristics of the test clearly reflect current models of language use and language ability, as these are specified in the design features of the test and realized in actual test tasks. The measurement characteristics of the test have been thoroughly investigated using the most sophisticated and current approaches available. Further-
more, the test developer has a well-defined program for on-going research and development, aimed at continuing to improve the qualities of the test.

References


Cummins, J. (1983) ‘Language Proficiency and Academic Achievement,’ In


Dept. of TESL–Applied Linguistics
UCLA
Los Angeles, CA 90024
U. S. A.