Validating the revised Test of Spoken English against a criterion of communicative success

Donald E. Powers, Mary A. Schedl and Susan Wilson Leung *Educational Testing Service, Princeton, NJ* and **Frances A. Butler** *University of California, Los Angeles, CA*

A communicative competence orientation was taken to study the validity of test-score inferences derived from the revised Test of Spoken English (TSE). To implement the approach, a sample of undergraduate students, primarily native speakers of English, provided a variety of reactions to, and judgements of, the test responses of a sample of TSE examinees. The TSE scores of these examinees, previously determined by official TSE raters, spanned the full range of TSE score levels. Undergraduate students were selected as 'evaluators' because they, more than most other groups, are likely to interact with TSE examinees, many of whom become teaching assistants.

Student evaluations were captured by devising and administering a secondary listening test (SLT) to assess students' understanding of TSE examinees' speech, as represented by their taped responses to tasks on the TSE. The objective was to determine the degree to which official TSE scores are predictive of listeners' ability to understand the messages conveyed by TSE examinees. Analyses revealed a strong association between TSE score levels and the judgements, reactions and understanding of listeners. This finding applied to all TSE tasks and to nearly all of the several different kinds of evaluations made by listeners. Along with other information, the evidence gathered here should help the TSE program meet professional standards for test validation. The procedures may also prove useful in future test-development efforts as a way of determining the difficulty of speaking tasks (and possibly writing tasks).

I Introduction

The Test of Spoken English (TSE) is a measure of the oral language proficiency of non-native speakers of English. Its primary intended uses are

1) screening graduate teaching assistants; and

Address for correspondence: Donald E. Powers, Principal Research Scientist, Educational Testing Service, Mail Stop 17-R, Rosedale Road, Princeton, NJ 08541, USA; e-mail: dpowers@ets.org

Language Testing 1999 16 (4) 399–425 0265-5322(99)LT179OA

2) certifying the speaking ability of medical and allied health professionals.

The original items in the TSE were selected from a variety of potential item types, largely on the basis of their correlations with the total score on the Foreign Service Institute (FSI) interview. The goal was to include items having high correlations with FSI scores and, to the extent possible, low correlations with TOEFL scores.

Recently, the TSE was revised so that it would better reflect current views of language acquisition and testing, specifically modern notions about communicative competence. In particular, three of the seven original sections of the test were deleted. These sections required examinees to read aloud, to complete sentences and to answer questions about a single picture. They were considered to be the least communicatively oriented portions of the test and, thus, less widely accepted as measuring oral proficiency. More communicative rationales for developing the test and for rating speech samples were also advanced. Instead of providing subscores for fluency, pronunciation and grammatical accuracy as did the original TSE, the revised TSE now reports scores on a scale of communicative language ability. This new scale takes into account the effectiveness of communication resulting from a number of linguistic, sociolinguistic, discourse and strategic competencies.

Because context is central to the construct of communicative competence, the revised TSE uses tasks that are specified in terms of seven contextual characteristics: test interviewer, audience, setting, topic, purpose, function and visuals. The purpose of these features is to provide appropriate and rich sociolinguistic and discourse features, in terms of task characteristics, and to enable examinees to engage their communicative language ability in responding to test tasks (Henning *et al.*, 1995). As a result of these changes, the revision of the TSE is now held by Educational Testing Service (ETS) to be:

> a test of speaking ability designed to evaluate the oral language proficiency of non-native speakers of English who are at or beyond the postsecondary level of education . . . [When] used in conjunction with other measures, it can help provide an indication of the examinee's ability to successfully communicate in English in an academic or professional setting.

The construct underlying the revised TSE is said to be the 'ability to accomplish specific language tasks comprehensibly, accurately, coherently, and appropriately with respect to specific interlocutor/audience, topic, and purpose' (Educational Testing Service, 1994: 1).

This conception is more or less consistent with thoughtful explications of communicative competence that have been provided by,

among others, Bachman (1990), Canale and Swain (1980), Chapelle *et al.* (1997), Douglas and Smith (1997), Duran *et al.* (1985), Henning and Cascallar (1992) and Stansfield (1986), each of whom has contributed to a better grasp of the concept. However, as Douglas and Smith (1997) have suggested, even though the term 'communicative competence' has been used for three decades, the concept is still not well understood.

The study described here focuses on what we have termed 'communicative success', defined as the ability of listeners to respond correctly, appropriately or positively to a speaker's message. Our interest was in determining the degree to which examinee performance on the TSE is related to the communicative success of those who listen to them.

1 Previous research on TSE

In their study of a prototype version of the revised TSE, Henning *et al.* (1995) found the correlation between scores from the old and the new versions of the test to be .83, suggesting that both versions tap similar constructs. (This high correlation does not, however, tell us what this construct is.) Moreover, the investigators found that performance on both the old and the new versions of the measure correlated strongly ($r = .75$ and $r = .82$, respectively) with an oral language interview which, according to the authors, is recognized by the Foreign Service Institute, the Interagency Roundtable and the American Council on the Teaching of Foreign Languages (ACTFL). This evidence is also consistent with the view that the two versions are measuring constructs that are very highly related.

In light of the strong relationship between scores from the old TSE and the revised test, much of the information supporting the validity of the former measure still has considerable relevance for the new test. For the original measure, Clark and Swinton (1980) found that the relationship between TSE scores and performance on the Foreign Service Institute's oral proficiency interview was very strong ($r = .79$ for 60 foreign teaching assistants). They also observed moderately high ($r = .51$) correlations between instructors' TSE scores and student assessments of instructors' ability to manage a variety of situations involving language skills, including the ability to answer student questions. (This more modest relationship may reflect the fact that students considered pedagogical skill as well as oral proficiency when making their ratings.) TSE scores also related strongly ($r = -.68$) to students' ratings of the degree to which an instructor's pronunciation interfered with student comprehension: the higher the TSE score, the less the interference due to pronunciation. More recently, Stricker

(1997) computed a correlation of .38 between TSE scores (from the actual revised test) and student ratings of teaching assistants' ability to communicate in English.

TSE validation research has also been conducted in non-academic settings. For instance, Powers and Stansfield (1983) showed that performance on the original version of the Test of Spoken English was reasonably strongly related to health professionals' judgements of whether individual speakers had sufficient oral language skills to function successfully in various health professions.[1]

The studies mentioned above suggest two major approaches to validating tests of oral language proficiency:

1) Relating test scores to performance on other direct measures of speaking ability, such as oral proficiency interviews. The oral proficiency interview is a valued criterion in part because the examiner is able to interact directly with the examinee, probing as needed for evidence of proficiency.
2) Relating test performance to other ratings, such as student assessments of an instructor's ability to handle specific language situations typically encountered in academic settings and to communicate in English overall. Such ratings are valued because they are indicative of the utility of a test in particular settings.

Each of these approaches has its particular strengths and limitations. In neither case is the criterion perfect, that is, free from all irrelevant influences. For example, an oral interviewer may be swayed by the personal characteristics of a speaker, and (student) raters may be influenced by other, non-communicative aspects of the context in which they make their ratings (for example, by the grades they receive from the person they are rating!).

2 A communicative competence orientation to validation

In addition to using the two approaches mentioned above, inferences from tests of oral proficiency might also be validated according to how well listeners are able to correctly understand and act upon a speaker's intended message; for example, by complying with the speaker's directions or by revising a document according to the speaker's instructions. Along these lines, the TSE Committee has suggested that when evaluating TSE performances, raters should consider the success with which the message has been conveyed (Douglas and Smith, 1997). The Committee has, however, taken a realistic view

[1]Because logistic regression was used, no product–moment correlations were reported. However, inspection of scatterplots suggests that the correlations were probably in the .50s.

of the promise of this approach, acknowledging the many, possibly uncontrollable, differences among listeners. Even native speakers vary a great deal with respect to their listening skills.

We believe that the approach we have taken is responsive to McNamara's (1997) concern that second language performance assessments have focused more on *language* assessment than on *communication*, which includes a social dimension. An objective of our approach was to tap this social dimension by allowing both the speaker and the listener to participate in the construction of meaning. In our study, the primary role of native-speaking listeners was as audience – to understand TSE speakers' messages. In carrying out this role, listeners were required to focus on the meaning of the message, rather than on discrete language skills. In this sense, then, listeners can be said to have interacted with the speaker in the active co-construction of meaning. This relationship reflects the revised TSE's emphasis on communicative success.

The approach that we describe below – that is, enlisting native-English-speaking undergraduate students as listeners/evaluators – is also supported by research: It appears that extensive experience in evaluating the speech of non-native speakers is not necessarily a prerequisite for rendering accurate judgements of non-native speakers (Mattran, 1977).

To implement the approach, we drew on research conducted under the rubric of referential or 'informing' communication, which provided a useful paradigm for the current study. Dickson and Patterson (1981) characterized referential communication as attempting to communicate to another person about a target referent. Accuracy is determined according to how well the listener understands the communication, as evidenced by the listener's ability to identify the referent among several alternatives or to reproduce it with some degree of fidelity. Among the kinds of task that have been studied are picture choosing, map directions and model building. In picture choosing, for example, the speaker directs the listener to choose one picture among several alternatives. The difficulty of the task can be increased or decreased by varying the demands placed on both the speaker and the listener. For instance, a map-directions task can be made more difficult by increasing the number of details on the map or the complexity of the route to be followed.

Yule (1997) has recently provided a comprehensive summary of research on referential communication. Here, however, we will describe only a single study – one by Lieb-Brilhart (1965) – to illustrate the approach. Lieb-Brilhart enlisted college undergraduates and asked them to describe preassigned geometric designs to a larger audience, whose members were asked to reproduce the designs as

described. Speakers were evaluated according to how well listeners were able to duplicate the designs that were described, and listeners in turn were judged by how well they reproduced the designs portrayed by each speaker. The referential communication paradigm has also been used to study interactive communication, when listeners are able to request additional information or clarification from a speaker.

We thought that this design might prove useful as a paradigm for test validation. The approach we envisioned would also, we believed, be viewed favourably by some language testers (Fulcher, 1987; Fulcher, 1996; Upshur and Turner, 1995), who have decried current rating scales for their 'lack of any empirical underpinning' (Fulcher, 1996: 208). In response, some language testers have attempted either to devise more empirically-grounded scales (Upshur and Turner, 1995) or to empirically validate existing second language proficiency descriptors (Butler and Stevens, 1998).

3 Objectives

The aim of the present study was to validate TSE score interpretations using a design that employed a communicative competence orientation, as we understand this notion. The goal was to marshal additional evidence of the validity of the revised TSE for its intended purposes. The specific research questions of interest were the following:

1) What is the relationship of an examinee's (a non-native speaker's) performance on the TSE to a listener's ability to understand the descriptions, information, directions (and so on) that the examinee is asked to communicate?
2) Is this relationship stronger for some TSE items than for others?
3) To what extent does this relationship depend on the listening proficiency, or other characteristics, of listeners?

II Method

To accomplish these objectives, we developed three separate instruments, administered them to a sample of undergraduate students, and analysed the resulting data as described below.

1 Instruments

The instruments used in the present study included a secondary listening test, a listening test and a background questionnaire. The secondary listening test was designed to collect listeners' evaluations of non-native speakers' responses. The two other instruments were used

primarily to address the fact that speakers' success may depend to considerable extent on factors that are beyond their control, such as the listening proficiency and other qualities of their audience. In order to take these factors into account, additional information was collected about the characteristics of listeners by administering a brief test of listening comprehension and and a background questionnaire.

a Secondary listening test: The validation criterion for the study was a specially developed secondary listening test (SLT). It was constructed in the following manner. For each of 12 TSE tasks (see Appendix 1) a set of 4–6 items was developed to assess listeners' understanding of TSE examinees. The items were of two general kinds:

1) ratings by listeners of:
   a. the amount of *effort required* to understand a speaker;
   b. the degree of *confidence that they had understood* a speaker;
   c. the extent to which a speaker's inability to communicate in English *interfered with understanding*.
2) questions that assessed *task fulfilment*, such as the degree to which listeners:
   a. were able to *identify* a speaker's purpose;
   b. could *follow* a speaker's instructions;
   c. found the speaker to be *persuasive*;
   d. could *reproduce* what the speaker said;
   e. exhibited *other kinds of evidence* that they understood the speaker.

Thus, the SLT was designed to reflect the more holistic aspects of the revised TSE scoring rubric, especially:

1) comprehensibility – i.e., the degree to which a listener is able to correctly identify the intended meaning of the speaker; and
2) effectiveness of communication – i.e., the degree to which an intended message was successfully conveyed to the listener.

Six alternate forms of the SLT were assembled as follows. Each form comprised exactly the same questions. Forms A through F differed only with respect to the particular taped TSE responses that accompanied them. Each of the 12 TSE speakers, used to represent the five TSE score levels (20, 30, 40, 50 and 60), is heard only once for each form, and each of a TSE examinee's responses is used only once across the six forms. Table 1 shows the scheme used to constitute the six SLT forms. For example, for Form A – TSE Task 1 – listeners first heard a response from a female TSE taker who had received a score of 20 on an actual administration of the test. Listeners

Table 1 Secondary Listening Test (SLT) design

TSE task	SLT form					
	A	B	C	D	E	F
1	20_F	30_M	40_F	50_F	60_M	60_{NM}
2	30_F	20_F*	50_M	40_F	60_{NF}	60_M
3	40_M	60_{NM}	20_F	60_F	50_M	30_F
4	50_F	60_F	30_M	60_{NM}	20_F*	40_F
5	60_M	40_F	60_{NF}	20_F*	30_M	50_F
6	60_{NF}	50_F	60_M	30_M	40_F	20_F*
7	40_F	60_M	20_F*	60_{NF}	30_F	50_M
8	50_M	60_{NF}	30_F	60_M	40_M	20_F
9	60_F	50_M	60_{NM}	30_F	20_F	40_M
10	30_M	20_F	40_M	50_M	60_F	60_{NF}
11	60_{NM}	40_M	60_F	20_F	50_F	30_M
12	30_F*	30_F	50_F	40_M	60_{NM}	60_F

Notes: Each table entry denotes the TSE score of the TSE speaker whose response is heard for the particular item. The subscripts F and M denote female and male speakers, and the subscript N denotes a native speaker of English. 60_{NF}, for example, denotes that the response is from a female native speaker who received a TSE score of 60; * Because of difficulty in identifying a male speaker at the TSE = 20 level, a female speaker was used as a substitute. This speaker had received a score of 30 on TSE question 12.

were then asked to answer a set of five questions (1.1–1.5) on the SLT. Next, Form A listeners heard a response to TSE Task 2 from a 30-level female TSE taker and were asked to answer a set of four questions (2.1–2.4) assessing their understanding of this speaker's response. This pattern continued until Form A listeners had heard 12 different TSE speakers – one male and one female representing each of the five TSE score levels (20, 30, 40, 50 and 60) and two native speakers of English, one male and one female.[2] This last provision was thought to be useful for establishing a baseline for comparing the primary results based on non-native speakers. All TSE responses were extracted from test protocols and assembled as SLT forms by a professional recording studio to ensure high quality reproduction.

For each of the six alternate forms, the order in which speakers were presented (again, see Table 1) was counterbalanced to control for any order effects (for example, of having heard a weak speaker before a strong one, and vice versa). Each listener heard a full range of responses (in terms of quality), but because a carry-over effect was possible, no listener heard more than one response for any TSE

[2]An exception was that no TSE 20-level male speaker could be identified, so a female test taker was used instead at this level. In addition, this 20-level female speaker had obtained a rating of 30 for TSE task 12.

task. (By carry-over effect, we mean the likelihood that listeners' success would be greater on later trials as the result of having heard an earlier speaker attempt the same task, regardless of how proficient the earlier speaker was on the task.)

It should be noted that the final version of the SLT (available from the first author) is the result of several iterations and try-outs, beginning with the development of a variety of alternative items and formats, which were administered initially to four ETS staff members to get preliminary reactions and suggestions for improvement. Each of the project investigators also tried the items. On the basis of these try-outs, a revised set of items was assembled, and procedures for administration were developed. These items and procedures were pilot-tested on a small number of students from a local college. The objective at this stage was to refine administration procedures (such as, how many times tapes should be played) and to identify any ambiguities in directions and items. The result was the deletion of several questions and the rewording of some directions.

Next, three of the six alternate forms of the SLT were administered to approximately 60 student volunteers at a local college. The data generated in this pretesting were formally analysed to determine the extent to which the various SLT questions distinguished among TSE score levels; that is, the degree to which the questions indicated that listeners had an easier time and were more successful in understanding speakers who scored higher on the TSE than those who scored lower. This strategy parallels that used in traditional test development practice: pretesting serves to identify the most promising test items for future operational administration.

b Listening test: In order to gauge the listening skills of study participants, a brief, 10-item listening comprehension test was assembled from items in CTB/McGraw-Hill's Listening Test (CTB/McGraw-Hill, 1985). This test is described in the test administration manual as a measure of the ability to follow directions and interpret connected discourse. The level used here (Level 6) is appropriate for upper-division secondary-school students and for first-year college students. Its administration requires test takers to listen to relatively brief passages and to answer questions associated with each one.

For this study, we selected three recorded passages and 10 associated comprehension questions, most of which required the recall of specific details from a passage. This abbreviated measure was used to classify study participants, albeit crudely, according to their listening skills. A concern, however, was that the test might prove relatively easy for most of the study participants, thus serving less as an indicator of listening skills than of participants' motivation to attend

to the demands of the study. Nevertheless, it was seen as a potentially useful indicator of a trait upon which the relationship between speakers' ability and listeners' understanding might depend.

c Background questionnaire: Building on the research of Rubin (1992) and others, a brief questionnaire (available from the first author) was developed to obtain additional information that might relate to study participants' understanding of non-native speakers. Information was requested about subjects' experience with and general attitudes about non-native speakers. Specific questions concerned the extent of participants' foreign-language study and travel, the nature and frequency of their contact with non-native speakers of English and the effects of any interactions with non-native speakers. The background questionnaire was pretested along with the SLT.

2 The sample

Mainly with the help of the TSE Committee, interested faculty were identified at the following institutions: Columbia University (Teacher's College), Iowa State University, Ohio State University, Northern Arizona University and West Virginia University. At each site, undergraduate students were recruited, and each was paid $20 for participating. Recruiters were asked to solicit a preponderance of first- and second-year undergraduates (as these students were most likely to have contact with non-native teaching assistants), to recruit students having a range of academic ability and to include non-native speakers in the sample. (Nonnative speakers were represented in the study sample because of their presence in the undergraduate population.)

The original design specified a total sample of about 450, composed of approximately 90 participants at each of the five sites. However, because recruiters experienced difficulty obtaining participants from one site, the actual distribution of participants across sites was 13, 72, 116, 75 and 162, respectively, with a total sample size of 438. The initial intention was to administer a different set of three SLT forms at each site, counterbalanced in such a way as to yield equal numbers of participants at each of the sites. Because of the recruiting difficulties at the one site, this ideal was not fully realized. Instead, at that site, only one form could be given; at another site, however, where more students were recruited, all six forms were administered, thus making up the shortfall. Forms were administered in the spring of 1997 to about 25 participants at a time.

As we had anticipated, study participants' performances on the CTB/McGraw Hill Listening Test were generally very strong. The

mean number correct was 8.4 of 10 ($sd = 1.5$). For the purposes of our study, we arbitrarily chose a score on the Listening Test so as to divide study participants into two roughly equal subgroups. Those who scored either 9 or 10 were classified as 'good' listeners (or motivated study participants) and those scoring 8 or lower were labelled as 'poor' listeners (or less motivated participants). On this basis, 54% of participants were classified as good listeners and 46% as poor. The dependability of the classification was modest, with $\Phi_\lambda = .57$ (Feldt and Brennan, 1989: 141).

With few exceptions, the sample was relatively homogeneous with respect to relevant background variables. Most participants were born in the United States (94%), had studied a foreign language in high school (91%), had at least some contact with non-native speakers in a typical week (89%), were White (79%) and were 18–21 years old (66%). The sample did, however, exhibit variation with respect to most other characteristics. For instance, whereas 36% reported that they had not travelled or lived outside the USA, a minority (13%) said they had done so for more than 36 weeks. A minority (16%) had not encountered any non-native instructors, but about as many (18%) had experienced five or more non-native instructors. Table 2 summarizes other characteristics of the study sample.

3 Data analysis

The aim of the analysis was to relate TSE score levels to speakers' communicative success, as we defined it. As is typical in most validity studies, a correlational approach seemed most appropriate. In our analyses, we used means as the unit of analysis. That is, for each SLT question, the mean performance of listeners was regressed on TSE score levels ($n = 6$). The responses of the two native speakers were treated as a higher TSE score level: in effect a 70 on a hypothetically extended TSE score scale. This strategy seemed reasonable, as the native speakers' responses were regarded by TSE raters to be *strong* 60-level responses. In the analysis, TSE score levels were coded as 20 = 1 to 60 N = 6.

Initially, we explored the fit of several functional relationships – linear, exponential and logarithmic – for each SLT question. The two particular non-linear methods seemed to be a reasonable selection from among many possibilities; also they are relatively familiar functions and therefore more interpretable than some others. Regressions were computed for *positive* responses as the independent variable; for example, the proportion of listeners who said they required little or no effort to understand the speaker. They were also computed for *negative* responses, such as the proportion who said they required a

Table 2 Study participants' relationships with non-native speakers

Type of contact	Percentage of participants
Frequent contact with nonnative speakers:*	
Friends/social acquaintances	29
Colleagues/business acquaintances	20
Teachers/teaching assistants	35
*Infrequent** contact with nonnative speakers:*	
Friends/social acquaintances	49
Colleagues/business acquaintances	63
Teachers/teaching assistants	42
Number of nonnative contacts during a typical week:	
None	10
1–2	44
3–5	25
More than 5	20
Length of typical encounter:	
Less than 1 minute	14
About 1–10 minutes	40
About 10–60 minutes	31
One hour or more	12
No contact	3
Number of courses taken in which instructor was a nonnative speaker:	
None	16
1–2	39
3–5	27
More than 5	18
Number of occasions on which final grade was hurt because instructor was a nonnative speaker:	
None	62
1–2 occasions	36
3–5 occasions	2
More than 5 occasions	0

Notes: Total $N = 438$; *ns* for individual questions ranged from 429 to 438; * 4 or 5 on a 5-point scale ranging from 'very frequent/daily or almost daily' to 'very infrequent/several times a year or less'; ** 1 or 2 on the 5-point frequency scale.

great deal or an extraordinary amount of effort. Regression analyses were also run both for 'good' listeners and for 'poor' ones, as determined by their scores on the CTB/McGraw Hill Listening Test. The results of these exploratory analyses suggested that a linear model generally provided the best fit for positive responses, and a non-linear model (in particular, a logarithmic function) was generally best for negative responses. These results held for both good and poor listeners. Thus, the results reported below are based on these two fits.

Finally, correlations were computed between performance on the SLT and the background information that was gathered.

III Results

Results are presented here for several analyses, using both descriptive statistics and regression methods. Results are also provided in terms of both SLT question types and TSE tasks. More detailed results are available in Powers *et al.* (1999).

1 Descriptive analyses

Table 3 illustrates the data that were collected for one TSE task ('Choose a place on a map and recommend reasons to visit it.') and for the SLT items associated with this task (1.1–1.5). (For this illustration we have presented data for the *total* sample, not for good and poor listeners separately, as is done for most of the analyses reported

Table 3 Illustrative results for one TSE task

TSE task/SLT item	TSE score level					
	20	30	40	50	60	60N*
				Per cent		
1. *Map: reasons to visit*						
1.1 Purpose (per cent correct)	10	38	88	89	73	95
1.2 Effort						
Little[a]	1	36	45	38	63	93
Much[b]	83	14	4	10	3	1
1.3 Confidence						
Much[c]	9	64	55	42	56	79
Little[d]	71	14	9	18	22	13
1.4 Persuasiveness						
Much[e]	0	1	1	11	22	28
Little[f]	93	61	46	13	24	13
1.5 Interference						
Little[g]	1	61	72	51	69	99
Much[h]	94	13	10	24	8	0

Notes:
[a] Hardly any or a limited amount;
[b] A great deal, an extraordinary amount, or couldn't understand at all;
[c] Quite certain/confident or extremely certain/confident;
[d] Quite uncertain, extremely uncertain, or couldn't understand at all;
[e] Very or extremely;
[f] Slightly, not at all, or couldn't understand the speaker at all;
[g] Interfered slightly or did not interfere at all;
[h] Interfered considerably or interfered completely;
* This level is for a native speaker of English.

below.) Table entries for row 1.1 are the percentages of study participants who successfully identified the speaker's purpose as 'giving directions'. Approximately 75 participants heard a 20-level TSE speaker, of whom 10% were able to correctly identify the speaker's purpose; of the approximately 75 listeners who heard a native speaker (60N), fully 95% could correctly specify what the speaker was trying to accomplish. Further, only 1% of those who heard the 20-level speaker said that they required little effort (hardly any or a limited amount) to understand this TSE examinee, while 83% said they needed much effort (either a great deal, an extraordinary amount, or they could not understand the speaker at all). Comparable percentages for the native speaker were 93% (little effort) and 1% (much effort). This information for each TSE task and for each SLT question is available from the first author.

2 Regression analyses

Figure 1 shows for another TSE task the most frequently observed relation between TSE score level and percentage of *positive* responses by listeners. In the example shown, the TSE task to which speakers had responded, after they had examined a sequence of six pictures, was 'Tell me the story that the pictures show.' The corresponding SLT question (4.3) asked listeners to indicate 'How certain/confident are you that you understood the speaker's reconstruction of the story?'

The relationship shown in Figure 1 is typical in the following ways:

- For both good and poor listeners there was a strong relationship between positive responses (in this case, confidence that listeners understood a speaker) and speakers' TSE scores.

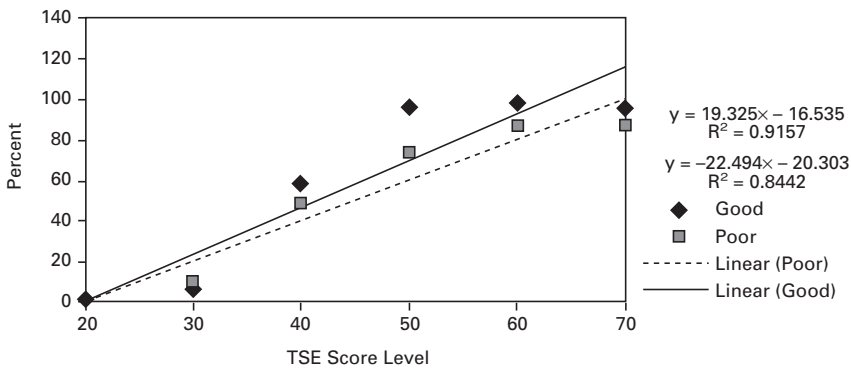


Figure 1 Regression of mean percentage of correct (or *positive*) responses on TSE score levels for good and poor listeners

- As suggested by a steeper slope, the relationship was slightly stronger for good listeners than for poor ones.
- Good listeners performed better than poor listeners (in this case, were more confident that they understood TSE speakers) regardless of the speaker's TSE score.
- The estimation of listeners' performance (that is, confidence) from TSE score levels was quite accurate, as evidenced by the high R^2 values for both good and poor listeners.

For all other positively-scored responses for each SLT question, information about the relationship between listeners' mean performance and TSE score level is available from the first author.

Figure 2 depicts for another TSE task the most frequently observed relationship for *negative* responses. This TSE task asked speakers to 'Discuss what [the information given in a graph] might mean for the future.' And SLT question 11.2 asked 'How much effort did you need to understand the speaker?'

Each of the characteristics of the typical relationship for positive responses (as exemplified in Figure 1) also applies, with one exception, to the most frequent relationship for negative responses. In contrast to the differential relationships found for positive responses, the relationships were equally strong (that is, the slopes were comparable in magnitude) for good and poor listeners. Again, as for positive responses though, relationships between TSE score levels and listeners' responses were strong. Additionally, just as good listeners were more likely to give *positive* responses, poor listeners were more likely than good listeners to give *negative* responses for each TSE speaker, regardless of TSE score level. Information about all other relationships for negative responses is available from the first author.

Table 4 summarizes the detailed results of the regression analyses.

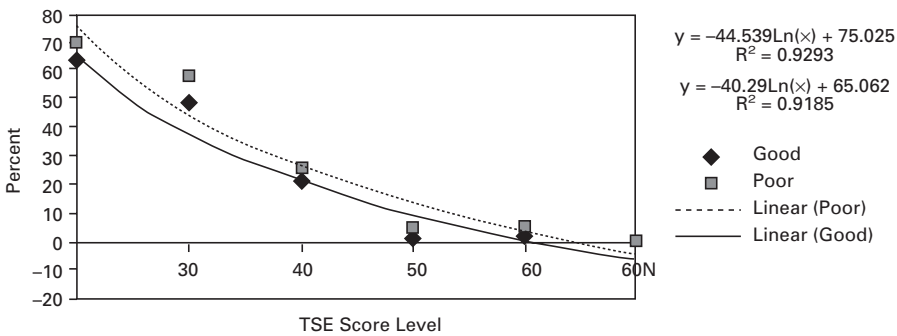


Figure 2 Regression of mean percentage of incorrect (or *negative*) responses on TSE score levels for good and poor listeners

Table 4 Comparison of difficulty and discrimination for good and poor listeners

Statistic	Greater for		*z*
	Good listeners	Poor listeners	
Positive responses (high confidence, low effort, etc.):			
Slope (strength of relationship)	44	15	3.65*
Difficulty	12	47	4.47*
Negative responses (low confidence, high effort etc.):			
Slope (strength of relationship)	25	20	0.60
Difficulty	6	39	4.77*

* $p < 001$, two-tailed sign test (Siegel, 1956: 68)

Characteristics of the relationships between TSE score levels and listeners' performance on SLT questions are compared for good listeners and poor ones. The table entries indicate the numbers of differences (and the *direction* of any difference) between good listeners and poor listeners for both positive and negative responses. Comparisons are made in terms of two statistics:

- the slopes of regression lines;
- the difficulty of each SLT item, as defined by the predicted percentage of listeners responding positively (or negatively) to the item at a TSE level of 40, the midpoint of the TSE scale.

Clearly, the main difference was a tendency for SLT items to be more difficult for poor listeners than for good ones. This finding applies to both positive and negative responses by listeners. A majority of items (47 vs. 12 for positive responses and 39 vs. 6 for negative responses) were more difficult for poor listeners than for good ones. Both of these tendencies were highly significant ($p < .001$) according to a statistical sign test (Siegel, 1956). For positive responses, the tendency for slopes to be greater for good listeners (44 vs. 15) was also significant ($p < .001$).

It should be noted again that while a linear function generally provided the best fit for positive responses, a curvilinear (logarithmic) function fit the negative responses somewhat better generally. The implication is that *negative* reactions to speakers may *decrease* more rapidly as speakers move up the TSE scale than *positive* reactions *increase*. Positive and negative reactions seem therefore not to be simple mirror images of one another. This finding may deserve further consideration and study to establish its psychological import.

With respect to the analyses by participants' listening scores, the intention was to determine any differential relationship between TSE

score level and listeners' understanding of speakers associated with listening ability. These analyses by listening proficiency can also be viewed as two independent analyses using two very similar samples. The similarity of results from this replication for good and poor listeners suggests that our findings are quite stable.

With respect to the other background information that was collected, no other variables were consistently related to listeners' performance on SLT questions.

3 SLT question types

Of interest also was the possibility that the various SLT item types might relate differently to TSE scores. Table 5 summarizes for each of five SLT question types information about the slopes and fits of the relationships between TSE level and responses to SLT items. On average, for each of the categories of SLT items, SLT listening performances are strongly related to TSE score levels. Only SLT items requiring listeners to identify a speaker's purpose ('Which of the following functions best describes the speaker's purpose?') were less discriminating than other kinds: the median slope (our discrimination index) was 12.2 for items of this type, compared with medians of 16.0 to 20.8 for the other four kinds.[3]

Table 5 Slopes (and R^2) of SLT performance on TSE score levels by SLT question type

SLT question type	*n*		Slope for		R^2 for	
			Positive responses	Negative responses	Positive responses	Negative responses
Purpose	7	Median	12.2	–	.64	–
		Range	5.8 to 18.2	–	.39 to .91	–
Effort	12	Median	20.8	–46.2	.86	.88
		Range	15.2 to 22.7	–27.5 to –60.4	.66 to .98	.71 to .97
Confidence	9	Median	16.0	–42.6	.87	.80
		Range	9.0 to 21.5	–8.7 to –57.1	.46 to .92	.57 to .96
Interference	12	Median	20.2	–50.3	.87	.89
		Range	14.0 to 23.1	–28.8 to –62.3	.66 to .97	.71 to .97
Task fulfilment	19	Median	16.1	–51.0	.87	.88
		Range	–8.3 to 22.3	–20.6 to –64.3	.21 to .98	.33 to .99

[3]The lower discrimination for 'purpose questions' may result in part from the context-relatedness of language functions. When responding to TSE tasks, examinees can safely assume that official TSE raters will know the context – that is, the particular questions that were asked – and respond accordingly, perhaps omitting important details provided in the TSE tasks. Participants in our study, however, did not have this context, especially for questions that required them to identify a speaker's purpose, which were always asked first – before the nature of the TSE task was gradually revealed by subsequent SLT questions.

Table 6, which shows the median intercorrelations among listeners' performances on SLT item types, reveals relationships that are consistent with expectations. Correlations were first computed among listeners' responses to SLT items for different TSE tasks and then averaged over the six SLT forms. The results show, for example, that the greater the effort required to understand a speaker, the lower listeners' confidence that they understood the speaker ($r = -.47$), and the greater listeners' perceptions that the speaker's inability to communicate in English interfered with understanding ($r = .60$). All question types correlated most strongly with ratings of interference.

4 TSE tasks

We felt also that listeners' performance on SLT items would shed light on the difficulty of TSE tasks. Table 7 contains the rank ordering of difficulty of TSE tasks by SLT question type for both positive and negative responses by listeners. Difficulty was defined here as the predicted proportion of positive (or negative) responses at a TSE score level of 40, that is, for an 'average' TSE speaker. Although the order of difficulty was not entirely consistent across SLT question types, it does appear that some TSE tasks are, according to our definition, more difficult than others. For instance, TSE Task 9 asks the speaker to define a term frequently used in the speaker's field for a listener who is not familiar with the speaker's field. This task is relatively difficult. TSE Task 12, on the other hand, appears to be relatively easy. This task requires the speaker to remind listeners about the details of an impending trip, and to inform them about changes that have been made to the itinerary. The fact that speakers are provided a chart showing the details and changes for the trip may be a factor in the relative ease of this task.

Some TSE tasks, we speculated, might be better reflections than others of the communicative success of speakers. Table 8 rank orders TSE tasks by their ability to distinguish among listeners' responses to each SLT question type; that is, by the degree to which TSE score level is associated with listeners' understanding. This association is

Table 6 Median intercorrelations among SLT item types

SLT question type	Effort	Confidence	Interference	Task fulfilment
Purpose	–.04	.39	–.51	.45
Effort		–.47	.60	–.33
Confidence			–.79	.60
Interference				–.72

Table 7 Rank order of difficulty of TSE tasks by SLT question type

TSE task	SLT question type								
	Purpose	Effort		Confidence		Interference		Task fulfilment	
		Low	High	Low	High	Low	High	Low	High
1	5	6.5	2.5	3	3.0	1.5	3.0	18.5	10
2	–	10.0	7.5	9	8.0	10.0	11.0	5.0	–
3	3	3.0	9.5	6	2.0	5.5	10.0	18.5, 14	13, 9
4	–	2.0	7.5	7	4.0	8.5	8.5	6, 8	7
5	–	8.0	9.5	5	5.5	8.5	6.0	2.0	–
6	7	6.5	11.0	–	–	5.5	7.0	7.0	3
7	2	1.0	2.5	–	–	1.5	2.0	13, 17	5, 3
8	4	10.0	6.0	2	5.5	4.0	4.0	3, 16	8
9	6	12.0	12.0	8	9.0	12.0	12.0	9.0	12
10	–	10.0	4.5	–	–	11.0	8.5	10, 1, 12	3
11	1	4.0	4.5	4	7.0	7.0	5.0	15, 11	6, 11
12	–	5.0	1.0	1	1.0	3.0	1.0	4.0	1

Notes: Easiest questions are rank ordered 1 for each SLT question, such that TSE tasks requiring little effort, for example, received a low ranking. When negative responses were considered (for example, the percentage of listeners indicating a high level of effort was required), the rankings have been reversed so that, again, a low ranking indicates a relatively easy task. Ties were treated in the conventional manner by summing ranks and dividing by the number of ties. Blank entries indicate that this type of SLT question was not asked for this TSE task. Multiple entries indicate that there was more than one SLT question for a TSE task.

indicated by the slope of the regression of SLT question performance on TSE score level. According to our criteria, TSE Tasks 4 and 6 seem somewhat better able than others to discriminate among speakers at various TSE levels. These tasks require TSE examinees to tell the story that a sequence of six pictures shows (Task 4) and to persuade a dry cleaner to clean the speaker's suit in less time than it normally takes (Task 6). TSE Task 1, on the other hand, is somewhat less discriminating than other TSE tasks. This task requires the speaker to give some reasons for recommending a place to visit.

IV Discussion

1 Summary

A secondary listening test (SLT) was constructed as a criterion against which to gauge the meaning of scores from the revised TSE. Stimuli for SLT questions were taped samples of examinee responses to the tasks posed by the revised TSE. The SLT was administered to several samples of undergraduate students to determine the degree to which speakers were successful in fulfilling the speech tasks posed

Table 8 Rank order of TSE tasks' discriminability by SLT question type

TSE Task	SLT Question Type								
	Purpose	Effort		Confidence		Interference		Task Fulfilment	
		Low	High	Low	High	Low	High	Low	High
1	2	12	8.0	9	8	12.0	10	17	10
2	–	8	5.0	5	2	8.0	4	4	–
3	3	4	1.5	7	4	5.0	1	18, 9	9, 3
4	–	1	3.0	2	1	1.0	2	12, 5	1
5	–	2	6.0	1	3	4.0	5	11	–
6	1	3	1.5	–	–	2.0	3	3	7
7	5	11	9.0	–	–	10.5	9	7, 15	6, 8
8	4	10	4.0	6	6	10.5	6	19, 14	2
9	6	7	11.0	3	5	3.0	8	2	5
10	–	9	10.0	–	–	6.0	7	6, 8, 16	11
11	7	6	7.0	4	7	7.0	11	13, 10	4, 12
12	–	5	12.0	8	9	9.0	12	1	13

Notes: The most discriminating questions were rank ordered 1. Ties were treated in the conventional manner by summing ranks and dividing by the number of ties. Blank entries indicate that this type of SLT question was not asked for this TSE task.

by TSE items. The communicative success of TSE examinees was defined according to listeners' reactions to speakers' messages; for example, the extent to which TSE speakers were judged to be successful in describing, directing, persuading or otherwise fulfilling TSE tasks. Success was measured operationally by listeners' performances on the secondary listening test, which posed several kinds of questions. For example, for each TSE task, listeners were asked to demonstrate their understanding of speakers' responses, to indicate the amount of effort required to understand them, and to record the degree of interference they experienced. One TSE item shows a map of a 'neighbouring town'. One of the TSE questions associated with the map asks: 'I would like to see a movie. Could you please give me directions from the bus station to the movie theater?' A corresponding task on the SLT required the listener to indicate the location on the map to which the speaker's directions, if followed, would lead. Several alternative forms of the SLT were developed, each based on a different order of the same speakers according to TSE score level. Each version was administered to a different, random sample of listeners.

For each TSE task, the relationship between a speaker's TSE score level and listeners' responses to SLT questions was computed. This served to assess the degree to which differences among TSE score

levels were associated with listeners' understanding. These relationships were computed for both 'good' and 'poor' listeners, as classified by their performances on a separate, standardized test of listening ability.

The results showed a strong association for the vast majority of relationships. Virtually without exception, the observed relationship was positive: higher TSE score levels were associated with greater understanding, less effort, more confidence and a higher likelihood that listeners could act appropriately in response to a speaker's message. For negative responses, the relationship was in the opposite direction. These associations were robust, inasmuch as they were observed when performances were analysed separately for good and poor listeners. The results also suggest that, according to listeners' performances on the SLT questions, not only are some TSE tasks more difficult than others, but some are more discriminating of speaking ability than others. The validity of our criterion – the SLT – as a measure of listeners' understanding was supported in that good listeners were significantly more likely than poor listeners to respond positively to each item.

2 Implications

The results of the study reveal the extent to which performance on each of several TSE tasks relates to communicative success, which we have defined here as the ability of listeners to respond correctly, appropriately or positively to a speaker's message. This outcome is consistent with claims made about the interpretation of TSE scores, and it represents one important step in accumulating the kinds of evidence needed to meet professional standards for test validation (AERA/APA/NCME, 1985). The results also point to some specific aspects (such as, effort, interference and confidence) that underlie both TSE scores and listeners' understanding, thus contributing to a better understanding of the test construct. The approach undertaken here is consistent, we believe, with Messick's (1989) notion of construct validation as the process of marshalling evidence in support of the meaning and use of test scores. Clearly, 'listeners' success' is a variable that ought to relate to TSE scores.

In addition to helping meet test standards, the information gathered here has utility, we believe, for better anchoring the meaning of performances on the TSE. By consulting *listeners*' performances associated with each TSE item score level, TSE users may gain a better grasp of what each *score* level signifies. The availability of a criterion measure – that is, a TSE-based secondary listening test – may also be welcomed by TSE users as a means for conducting local validation

studies. Local studies could be tailored by using samples of listeners appropriate to individual circumstances.

More generally, the results suggest the promise of a communicative competence approach to the validation of tests of productive language skills. The approach may also be applicable to tests of writing skill, and we suspect it may be easier to apply as well. A particular strength of the approach, we believe, is that it extends the base of 'validators' beyond the limited number of trained raters that are typically used to score tests like the TSE. Furthermore, it extends early research on the Test of Spoken English (Clark and Swinton, 1980; Powers and Stansfield, 1983) by providing an additional means of considering the views of undergraduate students, many of whom have become increasingly vocal about the quality of instruction provided by non-native teaching assistants, which is a major target of the TSE.

The procedures investigated here may also have utility for test-development efforts, as a method of evaluating any promising new kinds of items that may be considered for future versions of the TSE or for other measures of speaking. For example, the strength of the relationship between speakers' TSE scores and listeners' performances may be a useful index for selecting among potential new item types, with strong relationships suggesting likely prospects.

In addition, the methods we have used here may prove useful in scaling TSE and other language tasks according to difficulty level. As currently assigned by trained evaluators, holistic ratings of examinee performance on alternative TSE tasks are generally quite comparable across tasks, suggesting that all tasks are approximately equal in difficulty. However, Butler *et al.* (in press) have suggested that TSE tasks are *differentially* difficult, but that differences are masked when a generic scoring rubric is applied to all tasks. Our results lend support to the view that TSE tasks are in fact differentially difficult. Moreover, they suggest the possibility that, by indexing task performance to listeners' understanding of speakers, differences in task difficulty may be detectable.

With respect to scaling speaking tasks for difficulty, we note the success that some researchers have had in this regard. For instance, with Rasch modelling, Stansfield and Kenyon (1995) were able to estimate the differential difficulty of a variety of speaking tasks. Such tasks as 'describing a complex object in detail' and 'giving a professional talk' were among the most difficult; 'giving instructions' and 'introducing oneself' were among the easiest. Stansfield and Kenyon scaled tasks through ratings of the degree to which bilingual teachers should possess the level of ability implied by each task. Although need and difficulty do not necessarily co-occur, these researchers were able to quantify tasks, presumably according to their

difficulty. Coupled with these promising results, our findings suggest some avenues worth pursuing. Possibly, the kinds of judgements gathered here, especially when paired with more sophisticated methods of scaling than we have attempted, will constitute an even more direct and powerful basis for scaling the difficulty of speaking tasks than has been attempted to date.

Acknowledgements

The authors would like to thank the following people for their important contributions to the study reported here: members of the Test of Spoken English (TSE) Committee for advice and encouragement over the course of the study; Tony Ostrander and Evelyne Aguirre Patterson for advising us and (with Pat Stout) for providing TSE tapes; Peter Hagens and his staff at Hagens Recording Studio for their advice and production of audiotapes; Eleanore DeYoung, Ursula Ford, Gordon Hale, Karen Johnston and Ken Wilson for trying out the first version of our secondary listening test; Barry Yoder and students at the Philadelphia College of Bible for participating in pretesting of our instruments; Don Rubin, University of Georgia, for allowing us to borrow from his work in order to develop our background questionnaire; coordinators at participating institutions, who recruited participants and arranged for data collection and without whose help there would have been no study (the coordinators were: Barbara Plakans at Ohio State University; Joe Murphy, Helen Huntley and Jenny Yen at West Virginia University; Dan Douglas at Iowa State University; Bill Grabe, Joan Jamieson and Kristen Precht at Northern Arizona University; James Purpura and Jonathon Kim HyoSung Bidol at Columbia University); students at the participating institutions named above who provided the data for our study; Laura Jerry for analysing the data and producing the graphs for this report; Ruth Yoder for administrative assistance of all sorts; members of the Test of English as a Foreign Language Research Committee and the Committee of Examiners for suggestions regarding the design and reporting of the study and for their support of our efforts; and finally Dan Eignor, Barbara Suomi, Lyle Bachman and two anonymous reviewers for very helpful comments on an earlier draft.

V References

AERA/APA/NCME 1985: *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Bachman, L.F. 1990: *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Butler, F., Eignor, D., Jones, S., McNamara, T. and **Suomi, B.** In press: *A TOEFL 2000 framework for testing speaking*. Princeton, NJ: Educational Testing Service.

Butler, F.A. and **Stevens, R.** 1998: *Initial steps in the validation of the second language proficiency descriptors for public high schools, colleges, and universities in California*. Los Angeles, CA: University of California, Center for the Study of Evaluation, Graduate School of Education and Information Studies.

Canale, M. and **Swain M.** 1980: Theoretical basis of communicative approaches to second language teaching and testing. *Applied Linguistics* 1, 1–47.

Chapelle, C., Grabe, W. and **Berns, M.** 1997: *Communicative language proficiency: definition and implications for TOEFL-2000*. TOEFL Monograph Series MS-10. Princeton, NJ: Educational Testing Service.

Clark, J.L.D. and **Swinton, S.S.** 1980: *The Test of Spoken English as a measure of communicative ability in English-medium instructional settings*. TOEFL Research Report No. 7. Princeton, NJ: Educational Testing Service.

CTB/McGraw-Hill 1985: *Listening test: examiner's manual levels 1 through 6*. Monterey, CA: McGraw-Hill.

Dickson, W.P. and **Patterson, J.H.** 1981: Evaluating referential communication games for teaching speaking and listening skills. *Communication Education* 30, 11–21.

Douglas, D. and **Smith, J.** (with Schedl, M., Netten, G. and Miller, M.). 1997: *Theoretical underpinnings of the Test of Spoken English revision project*. ETS Research Memorandum RM-97–2. Princeton, NJ: Educational Testing Service.

Duran, R.P., Canale, M., Penfield, J., Stansfield, C.W. and **Liskin-Gasparro, J.E.** 1985: *TOEFL from a communicative viewpoint on language proficiency: a working paper*. TOEFL Research Report No. 17, ETS RR 85–8. Princeton, NJ: Educational Testing Service.

Educational Testing Service 1994: Test development documentation for the revised test of spoken English. Unpublished manuscript.

Feldt, L.S. and **Brennan, R.L.** 1989: Reliability. In Linn, R.L., editor, *Educational measurement*. 3rd edition. New York: Macmillan, 105–46.

Fulcher, G. 1987: Tests for oral performance: the need for data-based criteria. *ELT Journal* 41, 287–91.

Fulcher, G. 1996: Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing* 13, 208–38.

Henning, G. and **Cascallar, E.** 1992: *A preliminary study of the nature of communicative competence*. TOEFL Research Report No. 36, ETS RR 92–17. Princeton, NJ: Educational Testing Service.

Henning, G., Schedl, M. and **Suomi, B.K.** 1995: *Analysis of proposed revisions of the Test of Spoken English*. TOEFL Research Report No. 48, ETS RR 95–1. Princeton, NJ: Educational Testing Service.

Lieb-Brilhart, B. 1965: The relationship between some aspects of communicative speaking and communicative listening. *Journal of Communication* 15, 35–46.

Mattran, K.J. 1977: Native speaker reactions to speakers of ESL: implications for adult basic education oral English proficiency testing. *TESOL Quarterly* 11, 407–14.

McNamara, T.F. 1997. 'Interaction' in second language performance assessment: whose performance? *Applied Linguistics* 18, 446–66.

Messick, S. 1989: Validity. In Linn, R.L. editor, *Educational Measurement*. 3rd edition. New York: Macmillan, 13–104.

Powers, D.E., Schedl, M.A., Wilson Leung, S. and **Butler, F.** 1999. *Validating the revised Test of Spoken English against a criterion of communicative success*. TOEFL Research Report No. 63, ETS RR 99–5. Princeton, NJ: Educational Testing Service.

Powers, D.E. and **Stansfield, C.W.** 1983: *The Test of Spoken English as a measure of communicative competency in the health professions: validation and standard setting*. TOEFL Research Report No. 13. Princeton, NJ: Educational Testing Service.

Rubin, D.L. 1992: Nonlanguage factors affecting undergraduate judgments of non-native English-speaking teaching assistants. *Research in Higher Education* 33, 511–31.

Siegel, S. 1956: *Nonparametric statistics for the behavioral sciences*. New York: McGraw Hill.

Stansfield, C.W. 1986: *Toward communicative competence testing: proceedings of the second TOEFL invitational conference*. TOEFL Research Report No. 21. Princeton, NJ: Educational Testing Service.

Stansfield, C.W. and **Kenyon, D.M.** 1995: Comparing the scaling of speaking tasks by language teachers and by the ACTFL guidelines. In Cumming, A. and Berwick, R editors. *Validation in language testing*. Philadelphia: Multilingual Matters, 124–53.

Stricker, L.J. (1997.) *Using just noticeable differences to interpret Test of Spoken English scores*. TOEFL Research Report No. 58, ETS RR 97–4. Princeton, NJ: Educational Testing Service.

Upshur, J.A. and **Turner, C.E.** 1995: Constructing rating scales for second language tests. *ELT Journal* 49, 3–12.

Yule, G. 1997: *Referential communication tasks*. Mahwah, NJ: Lawrence Erlbaum.

Appendix 1 TSE tasks

PREPARING FOR THE TSE TEST

The TSE test is designed to measure proficiency in spoken English. Because spoken language proficiency can be achieved only after a relatively long period of study and much practice, an attempt to study English for the first time shortly before taking the test will not be very helpful.

To help you become familiar with the TSE test, several practice questions are provided below.

ON THE DAY OF THE TEST

On the day of the test, you will be given a test book and asked to listen to and read the general directions before you begin. It is a good idea to become familiar with the directions before the day of the test. The practice questions below are similar but not identical to questions you will find in the actual test. Therefore, responses to these practice questions may not be acceptable on an actual test. During the TSE test your responses will be recorded on tape. It might be helpful to record your practice responses on tape, then listen to hear how your speech actually sounds.

GENERAL DIRECTIONS

In the Test of Spoken English, you will be able to demonstrate how well you speak English. The test will last approximately 20 minutes. You will be asked questions by an interviewer on tape. The questions are printed in the test book and the time you will have to answer each one is printed in parentheses after each question. You are encouraged to answer the questions as completely as possible in the time allowed. While most of the questions on the test may not appear to be directly related to your academic or professional field, each question is designed to tell the raters about your oral language ability. The raters will evaluate how well you communicate in English.

As you speak, your voice will be recorded. Your score for the test will be based on your speech sample. Be sure to speak loudly enough for the machine to record clearly what you say. Do not stop your tape recorder at any time during the test unless you are told to do so by the test supervisor. If you have a problem with your tape recorder, notify the test supervisor immediately.

TSE PRACTICE QUESTIONS*

First, the interviewer will ask you three questions. These questions are for practice and will not be scored, but it is important that you answer them.

Sample questions:

What is the ID number on the cover of your test book? (10 seconds)

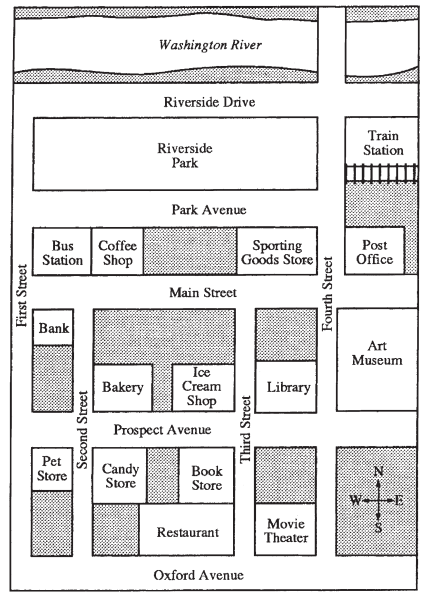
What is the weather like today? (10 seconds)

What are your plans for the rest of the day? (10 seconds)

Then the test will begin. Be sure to speak clearly and say as much as you can in responding to each question.

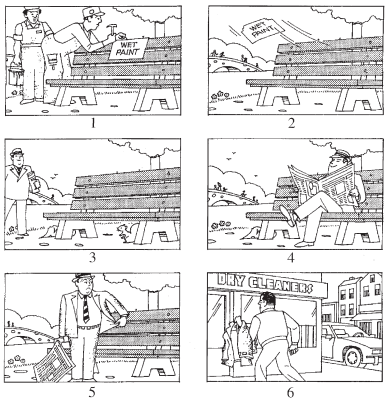
*Please note that the graphics used in the TSE practice questions are not the same size as those found in an actual test book.

Imagine that we are colleagues. The map below is of a neighboring town that you have suggested I visit. You will have 30 seconds to study the map. Then I'll ask you some questions about it.



1. Choose one place on the map that you think I should visit and give me some reasons why you recommend this place. (30 seconds)
2. I'd like to see a movie. Please give me directions from the bus station to the movie theater. (30 seconds)
3. One of your favorite movies is playing at the theater. Please tell me about the movie and why you like it. (60 seconds)

Now please look at the six pictures below. I'd like you to tell me the story that the pictures show, starting with picture number 1 and going through picture number 6. Please take one minute to look at the pictures and think about the story. Do not begin the story until you are told to do so.

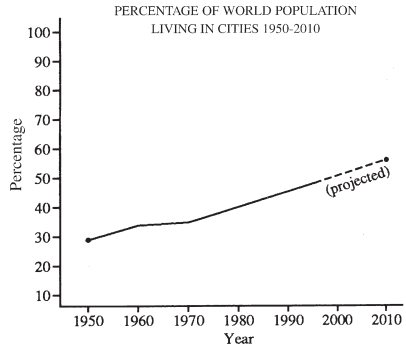


1 2 3 4 5 6

4. Tell me the story that the pictures show. (60 seconds)
5. What could the painters have done to prevent this? (30 seconds)
6. Imagine that this happens to you. After you have taken the suit to the dry cleaners, you find out that you need to wear the suit the next morning. The dry cleaning service usually takes two days. Call the dry cleaners and try to persuade them to have the suit ready later today. (45 seconds)
7. The man in the pictures is reading a newspaper. Both newspapers and television news programs can be good sources of information about current events. What do you think are the advantages and disadvantages of each of these sources? (60 seconds)

Now I'd like to hear your ideas about several topics. Be sure to say as much as you can in responding to each question. After I ask each question, you may take a few seconds to prepare your answer, and then begin speaking when you're ready.

8. Many people enjoy visiting zoos and seeing the animals. Other people believe that animals should not be taken from their natural surroundings and put into zoos. I'd like to know what you think about this issue. (60 seconds)
9. I'm not familiar with your field of study. Select a term used frequently in your field and define it for me. (60 seconds)
10. The graph below presents the actual and projected percentage of the world population living in cities from 1950 to 2010. Tell me about the information given in the graph. (60 seconds)



11. What might this information mean for the future? (45 seconds)
12. Now imagine that you are the president of the Forest City Historical Society. A trip to Washington, D.C. has been organized for the members of the society. At the last meeting you gave out a schedule for the trip, but there have been some changes. You must remind the members about the details of the trip and tell them about the changes indicated on the schedule. In your presentation do not just read the information printed, but present it as if you were talking to a group of people. You will have one minute to plan your presentation. Do not begin speaking until you are told to do so.

FOREST CITY HISTORICAL SOCIETY
TRIP TO WASHINGTON, D.C.

Date:	Saturday, April ~~8~~ 12
Transportation:	Chartered Bus
Depart:	8:00 ~~8:30~~ a.m. — Community Center parking lot
Itinerary:	10:30 a.m. — Guided Tour of White House
	12:30 p.m. — Lunch* - Rock Creek Park
	3:00 p.m. — National Museum of History and Technology (lecture - 4:00 p.m.)
	6:30 p.m. — Dinner - ~~Embassy Restaurant~~ Capital Inn Georgetown
Return:	10:00 p.m. (approximately)
Cost:	~~$20.00~~ $25.00 (excluding admissions and dinner)

* Bring your own

(90 seconds)