PHILIPPINE JOURNAL OF LINGUISTICS
Volume 34, Number 2 & Volume 35, Number 1
December 2003 & June 2004

TSE: PAST, PRESENT, AND FUTURE

John Miles
Educational Testing Service
Princeton, New Jersey, USA

This paper will summarize the development and status of the Test of Spoken English (TSE), discuss its strengths as an assessment of speaking proficiency, and indicate ways in which TSE will change to meet new demographic and technological needs as well as conceptual demands.

Originally, the TSE was intended to test international students desiring to become teaching assistants at North American universities. Because of U.S. immigration agency rulings in 1997 and decisions by various professional licensing boards, the test is taken by increasingly large numbers of medical and educational professionals. Revised in 1995 to be more communicative, the current TSE is a booklet-and tape-mediated test of general English speaking ability formulated to elicit a series of monologic speech samples. These recorded samples are evaluated at ETS by trained, certified and calibrated raters, who score each item holistically against the TSE Rating Scale. The test is statistically extremely reliable, with both high inter-rater reliability and high internal consistency between items.

Given changes in candidate demographics, in technology, and in linguistic concepts about speech, however, as well as the introduction of a speaking component in the new TOEFL test beginning July 2004, TSE will also change. Rather than a radical transformation that could be disturbing to test-takers, raters, and users (universities and licensing boards), a gradual evolution of the test is envisaged. Beginning in January 2003, for example, new item types will be pre-tested at successive administrations, and become operational by mid-year. Technological developments in the administration of TSE, such as digitization of test and sound files, are also expected.

The purpose of this presentation is to mark the changes that are occurring during 2003 in the Test of Spoken English (or TSE), a test produced, administered, and scored by Educational Testing Service (ETS), Princeton, New Jersey, USA. This test is well known in the Philippines, and yet there are many things about its past and present that are little known or misunderstood. One misconception that should be dealt with early and often is the false idea that the TSE has a pass-mark. As will be mentioned later, TSE scores place candidates in broad bands on a continuum of performance. It is those who use TSE scores who choose certain levels as cut-scores for their own purposes. ETS and the TSE staff and raters are not responsible for the pass-marks set by immigration services and professional organizations.

However, this presentation is not intended merely to deal with such misapprehensions. Rather, it is intended to shed light on the development of the TSE in the

past, the status of the TSE in the present, and the exciting modifications in test content, delivery, and scoring that are planned to occur over the coming months.

First, a little history. The Test of Spoken English has existed for over thirty years. In its original form, it was intended as a supplement to the TOEFL test and used written prompts to which a candidate had to respond orally into a tape recorder. The prompts involved sentence completion, response to simple questions, description of pictures, and so on. The test resembled closely the kind of exercises that would have been used in the language classrooms of the time.[1]

Between 1994 and 1995, a revised form of the TSE was designed by a committee of experts on the testing of speaking. It, too, represented what had become common procedure in second language classrooms in the United States where both theory and practice encouraged teaching for language proficiency or communicative competence. This new "Revised TSE" (the TSE in effect until December 2002) was a 12-item test based on a number of language "functions" that were considered essential to successful oral transfer of information. These functions included heavily sociolinguistic functions such as apologizing and recommending, and more obviously linguistic functions such as narrating, supporting opinion through argument or illustration, and hypothesizing or predicting. The 12-item TSE normally tested the ability of candidates to perform ten such functions.

This revised test was intended to be administered through booklet and cassette tape recorders, and the resulting speech samples to be scored by trained raters, almost exclusively ESL educators working together part-time on weekends at ETS. These raters were trained to rely on a detailed rubric that described the typical communicative performance of candidates at five levels or bands (characterized as 20-60) with reference to four principal competencies: functional competence, sociolinguistic competence, discourse competence, and linguistic competence. The Revised TSE was thus a criterion-referenced test that proved over the years to be a fair, valid, and statistically reliable way of determining how well someone could speak, always given the constraints of the method of delivery, the time allotted for preparation and speaking, the breadth of the bands of competence, and so on.

Because it was considered a fair, valid, and reliable indicator of speaking ability, the TSE quickly outgrew its use as a test for international students wanting to become teaching assistants in North American universities. Others saw the value of the test and began to use it. As was mentioned earlier, all of them were responsible for assigning the score that they desired their own test-takers to attain, their own "cut-score," using a score-setting kit designed by ETS. Most notably, in 1997, the Immigration and Naturalization Service of the United States began to accept performance on the TSE as a way for professionals (particularly those in the healthcare fields) to demonstrate their proficiency in speaking English. The INS set the bar at the 50 level, which contrasts with the score of 45 (indicating a level of communication that varies according to the function tested between "generally effective" and only "somewhat effective") set by most North American universities for international teaching assistants. The 50 level implies a steady level of high performance on all the tasks, whereas the 45 level allows for about half of a candidate's answers to be less consistently well spoken. The difference is important and explains why fewer candidates have over the years attained the popularly assumed pass-mark of 50 than might be anticipated by non-specialists.

Nevertheless, the TSE has grown as a test of speaking ability, being administered at test centers throughout the world and being used in some countries for certification of high-school English teachers and in several states in the United States for certification of potential bilingual teachers needing to demonstrate English speaking competence. Since test forms are designed alike, since speech samples are compared not to other candidate performances but to the identical criteria of the rating rubric, and since all candidates receive similar (if not identical) questions to respond to, the TSE has been considered extremely fair to all candidates. Its inter-item reliability, its inter-test form reliability, and its inter-rater reliability have been proved over and over. Indeed, the test has become so stable and reliable that it is statistically evident that twelve questions were no longer needed to ascertain the level of a candidate. This fact has allowed the rapid development of an enhanced, "new" TSE over the last few months and will enable its implementation during 2003, culminating in the inauguration of a 9-item TSE test beginning in September 2003.

The content of the traditional "Revised" TSE is well known by this audience: a map, a 6-picture sequence, a graph, a schedule, and a number of related or independent questions on issues that candidates were invited to discuss. The criticisms of this format are probably equally well known: the contexts seem frequently too general, the content seems often alienated from the experience of candidates, the tasks might sometimes be considered inappropriate (for example, some people do not have experience with reading maps or complex graphs), etc. From the viewpoint of academic experts, the test has become dated because it depends on a view of speaking as a separate skill and on a concept of the nature of linguistic functions that were considered no longer theoretically viable. The TSE Committee (a group of international experts on the teaching and testing of English) and the ETS staff who worked on TSE development have therefore agreed that changes were necessary.

Rather than a radical change in the test, however, it was decided to use the reliability of the Revised TSE by reducing the number of questions that contribute to candidates' scores and introducing new item types over a period of months in the form of pre-test questions. These experimental questions would be part of the monthly test administration but would not count against the score of candidates who were not accustomed to the new item types. Using a procedure known as Evidence Centered Design—a procedure used successfully over the past few years for the development of new tests at ETS—the Committee designed a number of new item types that would be developed by ETS staff and pre-tested with regular candidates and (according to customary TSE practice) pilot-tested with students in ESL programs in North American universities. From these item types, three new types would be selected to be part of a "new" TSE, together with the "traditional" item types that were considered most appropriate and helpful for evaluating the current candidate pool. The kinds of questions to be abandoned would be those considered least fair (like the map-reading kind), least discriminating (like the schedule kind), and those most obviously amenable to "canned" or pre-prepared answers (like the recommendation kind).

In addition, given the shift of the candidate pool from those with mostly academic to those with mostly professional preparation, it was decided that the context of some of the questions would change. Instead of a general or academic context, some of the questions—and all of the new item types—would have a workplace orientation. This would not require the use of particular business or health or other professional knowledge, but rather an understanding and awareness of the kind of English language used in a workplace environment and the ability to convey information in an acceptable form and manner.

Additionally, the new item types were designed to reflect the most recent understanding of language experts about the nature of speech as part of social interaction and would therefore involve an integration of what have been traditionally viewed as separate "skills." Thus, rather than just containing a spoken question with identical written words or a pictorial or visual stimulus as in the past, new item types could well contain aural stimuli, that is, spoken elements such as messages or conversations without any written support. Rather than testing listening, however, these kinds of items would test the integration of skills that is necessary in a conversational or presentational role. Indeed, all of the new items involve the playing of a role since candidates are called upon to imagine themselves functioning in English in the very workplace environment to which they almost all aspire.

Of the four new item types designed and developed at the time of this presentation, one involves no listening, one involves mainly listening, and two involve using both listening and visual cues. These item types can be seen and heard on the TOEFL/TSE Web site: www.toefl.org/tse or read in TSE publications such as the TSE Bulletin. Only the briefest introduction to these item types will be given in the following descriptions.

1. **Social Interaction**. Friendly exchange with coworkers is an essential part of the workplace environment, particularly important when it is necessary to go beyond the simple greeting and convey interest in and concern for others. In the setting of the question, candidates are given a specific context, a specific role, and a specific audience to address. The essential content of the remarks (a true exchange being naturally obviated by the use of an indirect test) is outlined in three bulleted phrases. These phrases, conveying three aspects of the topic of conversation that candidates are expected to elaborate on, are more formal than the language usually employed in social interaction situations in the workplace. So it is expected that candidates will use a generally more informal tone and register that is appropriate to the task proposed.
2. **Response to Voice Mail**. The telephone is an essential part of the modern workplace, and voice messages are an everyday occurrence for everyone: people listen to voice mail left for them when they are out or on the telephone already, and when they cannot reach others, they leave voice mail. This fact of modern life motivates this new item type, in which test candidates are asked to respond to a voice-mail message containing a complaint (or other message requiring a response) by leaving a voice message of their own. This kind of question requires test candidates to react empathetically to the person complaining (or asking for something) and to deal with the problem, need, request for information, or whatever. Again, candidates are given a context, a role, and an audience. They must choose the appropriate tone and register and respond with relevant content to the challenge posed by the caller.
3. **Progress Report**. This item type also uses the voice message as the mode of response used by candidates. However, the stimulus is not a voice message but a conversation between colleagues to which candidates (in their unspecified role as an employee or member of a business or faculty department) are witnesses. With the support of a simple visual (a list, flowchart, calendar, etc.), candidates are asked to make a voice-mail report to a supervisor about the status of the project that they heard their supposed colleagues discussing. Essentially, candidates are expected to add detail to the barebones visual and report on what has been accomplished and what remains to be done to complete the project. Again, there should be a difference between the more familiar

register and tone of the conversation to which candidates listen and the more formal register and tone appropriate to the voice-mail report to a superior.

4. **Conflicting Information.** In this item type, a voice message heard by candidates contains information that is different from the information on the same topic contained in a visual—a piece of "realia" (a business form, calendar, schedule, etc.). Candidates are asked to report the discrepancies and suggest ways to resolve the problems that present themselves in this conflicting information. Again, the context, role, and audience are specified, while the content of candidates' speech is specified by the differences between what is seen and what is heard. [After this presentation was delivered, it was decided that the fourth item type would not become a part of the 9-item TSE at this time.]

Obviously, these kinds of tasks require both time for candidates to think through a response and permission for candidates to take notes so as to organize what has been heard or given as basic content for the response. Both preparation time and note-taking are therefore provided—the second, at least, being a break with the practice of earlier forms of the TSE. Moreover, a full minute (often the time allowed by answering machines) is allowed for the response, though good candidates may not require a full minute to complete the task adequately.

Considerable though this change in test content is, it is not the only radical change that TSE will undergo during 2003. Traditionally, as was mentioned, the TSE has been administered in test centers using cassette recorders both to deliver the test and to capture candidates' speech samples. This is a cumbersome process fraught with the potential for problems: non-delivery of materials, loss of cassettes in the mails, poor recordings due to low level of recording or background noise (or both). It also involves a considerable time delay before rating can take place because of mailing times, as well as the need to bring together raters in one place to listen to cassettes and score the taped responses, one cassette after the other.

After much research, ETS has announced that, as of March 2003, the TSE will be available to be delivered by telephone from a central computer system. The process is called Interactive Voice Response or IVR, and several centers in the Philippines are gearing up to lead the way in this endeavor. Candidates will still need to use a test booklet, but they will hear and respond to the test by telephone. For convenience, test centers administering the TSE by the IVR system will generally provide headphones with built-in microphones for candidates.

The advantages are evident. First, the sound quality of the test will be immensely improved for candidates: most background noises, echoes, and the poor sound quality of the delivering cassette player will be eliminated. Second, the sound quality of the responses as captured by the central computer through the telephone will be considerably better than that captured by cassette recorders: again, background noises will be minimized and, thanks to computer technology, many otherwise inaudible responses will be able to be scored by raters listening by telephone or through the computer. Third, the time required to score responses will be lessened, since raters will not have to sit through the repetition of the prompts (or questions) or the preparation time, or try to fast-forward their cassette recorders: scoring using IVR will be more focused and more rapid.

Moreover, it is expected that because raters will not need to score the whole test of an individual candidate (a whole cassette) but only one or two responses of a candidate as provided by the computer over the telephone, IVR scoring will provide an even more accurate assessment of candidates' levels of proficiency than the traditional mode. Instead of just one or two raters, a multiplicity of raters will contribute to a candidate's overall score with both single and multiple ratings as required by the statistical model. Lead raters will be able to monitor the process from central computers and facilitate the whole rating process, which should inevitably be speedier. This means that, once the IVR system is fully implemented, test scores for the thousands of candidates at each administration of the TSE will be more quickly available to score-users and candidates.

It is therefore with a great deal of pleasure that I am able to announce these developments first to our colleagues in the Philippines, where the TSE has become so popular. And it is with anticipation and excitement that the staff at ETS looks forward to the implementation of a "new" TSE, apparently shorter but more integrated and inclusive, and a new method of delivering and scoring the test, technologically more advanced and psychometrically more advantageous.

---

[1] Since this was a presentation and not a scholarly paper, there are no footnoted quotations or learned references. However, those present at the conference will recognize that this written paper is different in form but not in content from the original presentation .