How ETS Scores the TOEIC® Speaking and Writing Test Responses

Philip Everson and Susan Hines

In order to understand ETS's approach to constructed response scoring for the TOEIC® Speaking and Writing tests, it is necessary to outline some challenges and assumptions.

- Performance, unlike responses to multiple choice questions, but like real life, must be evaluated over a continuum.

- Human raters have biases that can influence their behaviors when scoring.

- Rater behavior can be controlled.

- Multiple independent ratings will produce a more reliable picture of a test taker's proficiency than a single rater model.

Sources of Score Variability

There are several potential sources of variability in test scores including, but not limited to candidate behavior, form variability and rater behavior. The first, candidate behavior, is something that ETS cannot control. Many factors can contribute to candidate behavior on a test day: level of confidence, reasons for taking the test (e.g., job application, promotion, curiosity), health and legitimate changes in abilities. None of these variables are ones that ETS can influence. The other two sources of score variability are ones that ETS can and does control. Form variability is related to the level of difficulty across different test forms. If Test Form A is more difficult than Test Form B and a company is interested in hiring candidates who may have taken one form or the other, then it is problematic to compare candidates against one another because their test scores do not have the same meaning. ETS controls form-level variability through the creation and use of detailed test specifications and content reviews. In a perfect world, human raters would behave identically to each other when assigning scores. Score reliability depends on accurate and consistent behavior of the raters. However, this scenario does not represent the scoring environment in the real world. This paper describes how ETS addresses the challenges presented by the use of human raters for large-scale, high-stakes testing for the TOEIC Speaking and Writing tests.

The Online Scoring Network (OSN)

The scores that are produced by the TOEIC Speaking and Writing tests begin with the rating of the responses by trained raters using the ETS Online Scoring Network (OSN). Both the software platform through which rating occurs and the training and monitoring of the raters are essential to producing valid and reliable scores.

After a test taker completes a TOEIC Speaking test or TOEIC Writing test, his or her responses to test questions are sent via the Internet to ETS's patented secure scoring platform, OSN. Scoring for the TOEIC Speaking and Writing tests can best be summarized by these key design features:

- training and calibration of all raters supported within the scoring interface

- rater performance monitored as it is happening

- test responses distributed to raters anywhere in the world for rating

- the same criteria applied in the same way to all responses by each rater

- results of ratings collected quickly and efficiently

This scoring system provides the technical infrastructure needed to conduct scoring sessions that are data-driven, controlled, traceable and anonymous. The latter two features may seem contradictory, but they are described in more detail in the next paragraph.

The OSN system has been used with the computer-based TOEFL® CBT test since the mid 1990's. OSN's web-based design allows ratings to be assigned via the Internet over a secure connection to ETS. The OSN system is secure. Raters can only access the system when they have been scheduled to score. Raters are able to work from home, which means ETS can recruit large numbers of raters from across the United States and Canada. The OSN system stamps time and rater identification after each score is assigned, so in this sense the ratings are traceable. At the same time, the OSN system prevents any personal information about a test taker from being revealed to a rater while scoring, so in this sense the test taker is anonymous to the rater. This plays a key role in reinforcing accurate, fair, and consistent scoring.

Certification and Training of Qualified Raters

Raters for the TOEIC Speaking and Writing tests must be college graduates with experience teaching English as a second language or English as a foreign language at the high school, university, or adult learning levels. Raters are recruited through professional organizations, conferences and the ets.org Web site. Because TOEIC ratings take place via the OSN system, raters can be recruited from almost any geographical area in the United States and Canada.

Those who express interest in being TOEIC raters submit resumes that are reviewed by ETS. If the prospective rater has the appropriate education and experience, ETS gives him or her access to the Learn OSN TOEIC training Web-site. In training, raters learn about the TOEIC Writing test question types or the TOEIC Speaking test question types and how to apply the scoring rubrics to test takers' responses. Training materials include responses to several different versions of each question type. Training materials also include:

- detailed explanations of the purpose of each question type and how it contributes to the test taker's score

- benchmark responses - authentic responses that have been chosen by the ETS staff to represent each point on the scoring guide

- sample responses - authentic responses that raters-in-training use to practice rating; the prospective rater scores each sample and then checks the score against the rating and explanation given by the ETS staff

- written explanations of all scores assigned to benchmarks and samples

Prospective raters train on either speaking or writing question types. No rater rates for both tests. The rater pool is separate.

Prospective raters work on the training materials at their own pace. When they feel they have mastered the scoring guides for each of the question types in either the TOEIC Speaking or Writing test, they contact ETS and schedule a certification test. The certification test is a set of test taker responses with known scores. The trainee rates each response in the set without knowing the assigned scores. If the trainee's scores do not agree with the assigned scores, he or she must undergo more training before another opportunity is given to take a different certification test at a later date. If the trainee's scores agree with the assigned scores, the trainee qualifies to be a TOEIC Speaking or Writing test rater.

Daily Calibration

Even though every rater has passed a certification test, raters are required to pass a calibration test every time he/she logs into the OSN system. This calibration test assesses a rater's readiness to score for that specific scoring day. In the unusual case where a rater is unable to pass a second calibration test, he or she is dismissed from scoring for that day and asked to review training materials before his or her next scheduled scoring session. There are times when raters must take more than one calibration test. For example, a rater for the TOEIC Writing test may begin his or her scoring session by scoring question 4 (write a sentence based on a picture), but then be asked to score a different type of question later in the day, question 6 (respond to an e-mail). Because these represent different types of questions, a different calibration test that represents question-type 6 responses must be passed in order to proceed with the scoring session. OSN is programmed to prevent raters from accessing responses if the rater has not passed a calibration test.

Support and Monitoring During Scoring

During a scoring session, all raters have access to several OSN functions that provide support. First, benchmark responses, responses that represent the middle of each score level, are accessible at all times. All raters are required to listen to or read these benchmark responses before their scoring session begins and after they have returned from a break. Second, when necessary, the assessment specialists at ETS write special scoring instructions called topic notes that appear on the scoring screen for every rater to see during scoring.

To maintain ETS's high standards for score quality, every rater is assigned to a team. Each team includes up to eight raters and a scoring leader. The primary function of a scoring leader is to monitor rater behavior to ensure that every rater is accurately following the scoring guides. At any time, a scoring leader can stop a rater and discuss issues. Of course, at any time a rater can contact the scoring leader with questions during scoring, as well. Scoring leaders' contact information is available in OSN and all scoring leaders are required to contact every rater in their groups by phone throughout the scoring day. More detailed OSN functions are accessible to scoring leaders. The rater performance function allows scoring leaders to see scores as they are being assigned in real-time. Rater statistics are also available for a scoring leader. A representative sample of every question type is scored by two different raters for the purposes of collecting inter-rater reliability statistics (the percentage of two different raters assigning the exact same score to the same response). This functionality is a tool that can be used to identify raters who may need additional support during a scoring session. Scoring leaders receive face-to-face on-site training at ETS headquarters in Princeton, New Jersey and attend yearly on-site refresher trainings in addition to regular conference calls with ETS assessment specialists throughout the year.
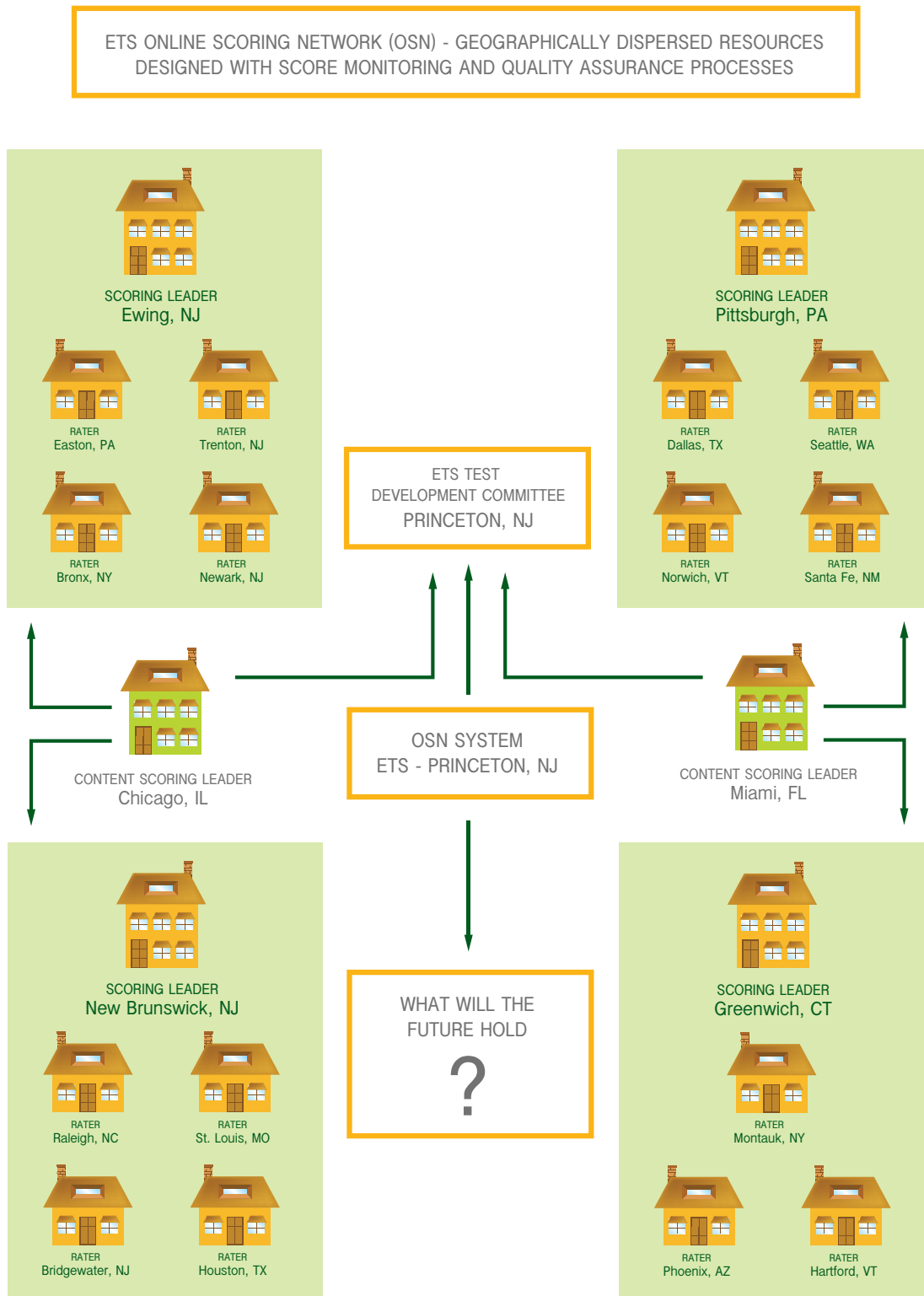
In cases of responses that are particularly difficult to hear, understand, or interpret, raters have the option of saying, "I don't know which score is best for this response" (This marks an important difference between this kind of scoring and scoring an interview test.). A scoring leader may also want additional guidance on a particular response that is anomalous in some way. In these cases, responses can be sent to the Deferral queue within OSN to be scored by a chief content scoring leader. The chief scoring leader creates and directs the scoring plan for the day (decides on the order of scoring, assigns scoring leader groups to different questions, etc.). Chief scoring leaders are identified by education, experience, scoring accuracy, and communication skills and often have similar qualifications to the ETS assessment specialist staff who write the test questions. Chief scoring leaders also provide mentoring and support to scoring leaders. This is in addition to their most important role, which is to alert ETS assessment specialists to high priority issues that may affect the accuracy of scores that are being assigned during a scoring session. The assessment specialists at ETS monitor every scoring session and rely on the chief scoring leaders to escalate important scoring issues. This close partnership is maintained with monthly phone conferences and regular communication throughout the week.

Both scoring leaders and chief scoring leaders submit end-of-day reports for their scoring session. This report is a record of the day's discussions with raters and provides a concrete reference for follow-up action that may be needed. ETS assessment specialists and chief scoring leaders work together to identify raters who may need additional mentoring, those who should be considered for promotion to scoring leader, and resolutions of any outstanding scoring questions at the end of the scoring day.

The OSN is distributed (see Figure 1). Because OSN sends test takers' responses to the raters over the Internet, TOEIC raters are not limited to any geographical area. This allows ETS to recruit qualified raters from across North America, or even the world in the future. By greatly expanding the pool of potential raters, OSN allows ETS to set very high standards and qualifications for those who are selected to be TOEIC raters. Additionally, because raters do not have to assemble at a single place at a designated time, ratings can be scheduled for any time, allowing for efficient return of the results to the test takers.

FIGURE 1

The ETS Online Scoring Network

ETS ONLINE SCORING NETWORK (OSN) - GEOGRAPHICALLY DISPERSED RESOURCES
DESIGNED WITH SCORE MONITORING AND QUALITY ASSURANCE PROCESSES

SCORING LEADER
Ewing, NJ

RATER
Easton, PA

RATER
Trenton, NJ

RATER
Bronx, NY

RATER
Newark, NJ

SCORING LEADER
Pittsburgh, PA

RATER
Dallas, TX

RATER
Seattle, WA

RATER
Norwich, VT

RATER
Santa Fe, NM

ETS TEST
DEVELOPMENT COMMITTEE
PRINCETON, NJ

CONTENT SCORING LEADER
Chicago, IL

OSN SYSTEM
ETS - PRINCETON, NJ

CONTENT SCORING LEADER
Miami, FL

SCORING LEADER
New Brunswick, NJ

RATER
Raleigh, NC

RATER
St. Louis, MO

RATER
Bridgewater, NJ

RATER
Houston, TX

WHAT WILL THE
FUTURE HOLD

?

SCORING LEADER
Greenwich, CT

RATER
Montauk, NY

RATER
Phoenix, AZ

RATER
Hartford, VT

### Multiple, Independent, Anonymous Ratings for Every Candidate

The distributed scoring design differs significantly from tests designed to be scored by a single rater's holistic impression of an entire test. Multiple, independent ratings increase the reliability of a final test score. First, the effects of an individual rater's behavior on the test taker's final score are minimized. Second, each independent rating contributes its full intended weight to the final score. When one rater scores an entire test, the first responses the rater scores tend to influence the ratings of later responses. This is called the halo effect. When independent ratings are given to each question, raters are not influenced by other responses from the same test taker.

The OSN supports independent rating. Because TOEIC Speaking and Writing tests are administered by computer and scored in OSN, each candidate response is a separate text or sound file. OSN takes all the test takers responses from an administration and organizes and separates them into folders by question-type. Raters work on only one type of question at a time. This means that for each test taker, one rater rates the response to Question 1, another, different, rater rates the response to Question 2, and so forth. Each rater does not know how the test taker performed on the rest of the test, so each rating is based on the response being evaluated, and only that response. The independence of each decision contributes to the overall fairness and reliability of the test results. Once the maximum number has been reached, OSN prevents a rater from scoring any more questions from that individual candidate. Rather than the final score representing one rater's impression of the candidate's overall performance, the final score for a TOEIC Speaking or Writing test represents the scores assigned by several different raters. A minimum of three different raters contribute to the final score of an individual TOEIC Speaking test and a minimum of three different raters contribute to the final score of an individual TOEIC Writing test. In actual practice, there is an average of 10 different raters who contribute to each test taker's final test score for a TOEIC Speaking and Writing test. OSN is designed to track the number of questions a rater has scored from an individual test taker. See Figure 2.

FIGURE 2.

How the Online Scoring Network supports independent rating

The OSN is anonymous. Raters do not know whose test they are scoring. They do not know the test taker's nationality, education, age, job, experience or any of the other factors that might bias human raters. They do not know why the person is taking the test or who will see the test scores. The raters are only interested in applying an accurate rating to each response.
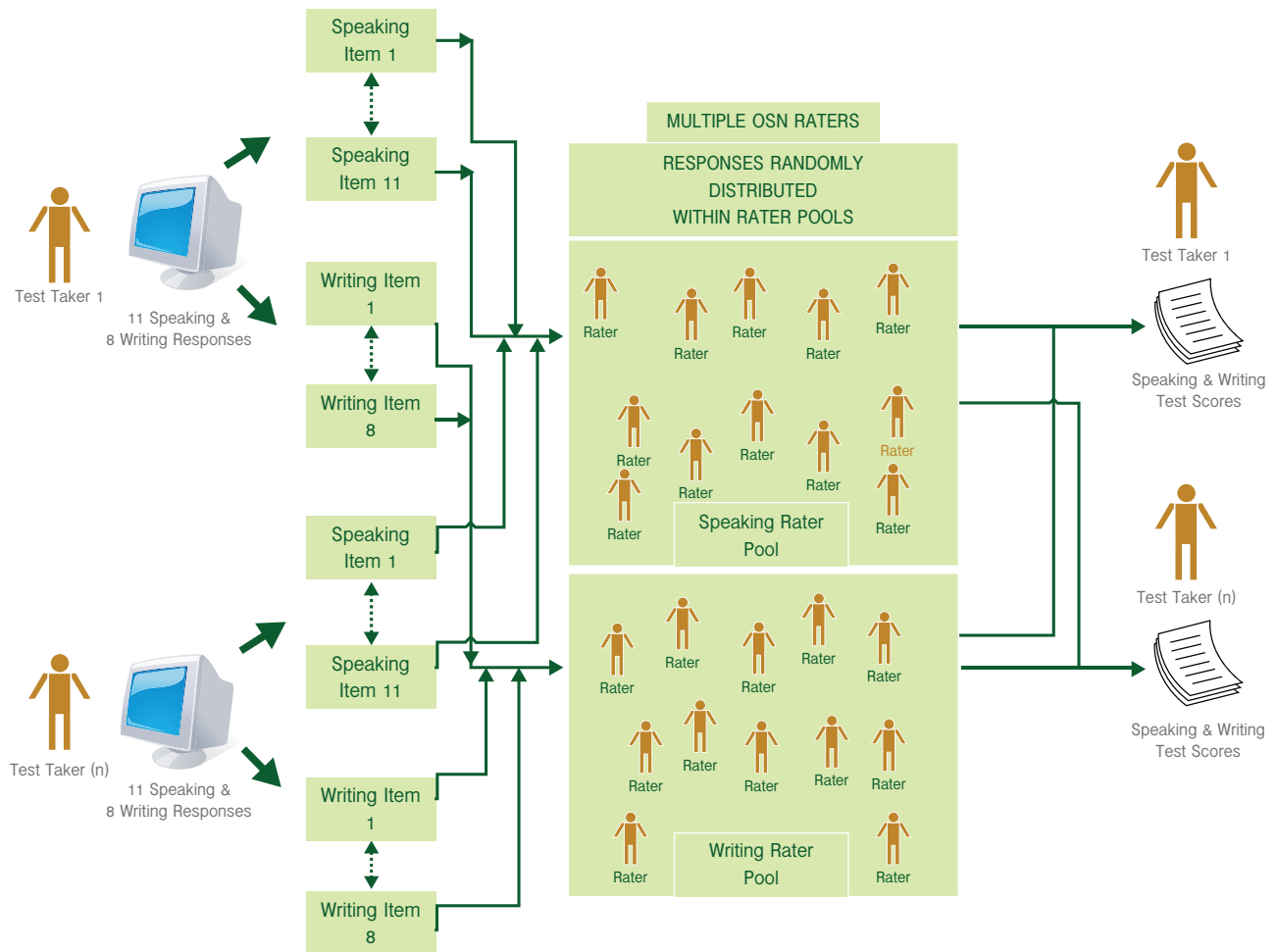
Correction of Discrepancies

There are times when a scoring discrepancy may occur. A scoring discrepancy for the TOEIC Speaking or Writing test occurs when there is a difference of two score levels between two raters' scores for the same response. For example, if Rater A assigns a score of 3 and Rater B assigns a score of 1 to the same response, OSN detects the discrepancy immediately and the response is automatically routed to an adjudication queue, where it is scored by scoring leadership. The adjudicator's score is the third and final rating for any discrepancies detected by OSN. Only scoring leadership can assign scores in the adjudication queue, so this final score is assigned by experienced raters. Because this detection of discrepancies is immediate, scoring leaders and/or chief scoring leaders can investigate the discrepancy and provide immediate feedback to raters who need additional mentoring.

Once the rating session is over, OSN collects all the ratings for each test taker. The raw ratings are then combined and weighted. The ratings on the most difficult of the items contribute more to a high score than do the ratings on the less difficult items. The combined raw ratings are converted to a scale score of 0-200. The scale scores are also divided into levels. The levels represent significant differences in performance on the tasks and are based on empirical observation of test takers' performance on the TOEIC Speaking and Writing tests.

Conclusion

Testing speaking and writing skills is valid, transparent, and necessary for today's global community. ETS's patented OSN platform:

- distributes test responses to raters anywhere in the world for rating

- supports training and calibration of all raters

- assures that each rater applies the same criteria in the same way to all responses

- enables monitoring of rater performance as it is happening

- collects the results of the ratings quickly and efficiently

Distributed, anonymous and monitored scoring means fairer and more accurate score results so that score users can make the best decisions for hiring, promotion, training and placement.