Compendium Study

Monitoring Individual Rater Performance for the *TOEIC*® Speaking and Writing Tests

Yanxuan Qu and Kathryn L. Ricker-Pedley

September 2013

Know English. Know Success.

One major issue for tests with constructed-response (CR) items is the reliability and accuracy of scoring. Responses to CR items are typically rated by trained human raters, which are subject to a variety of rater effects. For example, different raters may have different understandings of the scoring rubric (Saal, Downey, & Lahey, 1980); raters may be differentially stringent in scoring; raters may tend to use some score categories more often than others; or raters' rating behavior may drift over time due to fatigue or other factors (Fitzpatrick, Ercikan, & Yen, 1998; Hoskens & Wilson, 2001). The existence of rater effects will introduce measurement errors to test scores and thus will harm the usefulness of a test.

Despite the inherent scoring issue, tests with CR items are appealing in the sense of directly measuring productive skills that closely approximate tasks encountered in daily life. They also eliminate the possibility that test takers can answer correctly by guessing among multiple choices. For these reasons, tests with CR items are widely used by many large-scale testing programs in high-stakes tests. It is critical for every testing program using CR tests to enhance scoring consistency and accuracy by training or monitoring raters or by conducting statistical adjustments (Allalouf, 2007; Dunbar, Koretz, & Hoover, 1991). For all tests with CR items, training and monitoring raters is a continuous process that occurs throughout the whole scoring period.

The purpose of this paper is to describe procedures implemented by the *TOEIC*® Speaking and Writing tests for monitoring rater performance and enhancing overall scoring quality during and after each administration. The focus of this paper is on monitoring and improving raters' performance at the individual level so that trainers can provide more targeted training or retraining to raters for the TOEIC Speaking and Writing tests.

The following section introduces the current procedures developed to monitor overall and individual rater performance at the item level both during and after each administration. Future directions for monitoring rater performance for the TOEIC Speaking and Writing tests are also provided.

Current Procedures for Monitoring Rater Performance

Since December 2006, the TOEIC Speaking and Writing tests have been administered at Internet-based test centers in many countries all over the world. The tests are designed for non-native English speakers to measure their language production skills in daily life or in workplaces where English is required for communication. The two independent tests can be administered either together or separately. The speaking test has 11 tasks and takes about 20 minutes to complete, and the writing test has 8 tasks and takes about 1 hour to complete. All the items are CR format with varying numbers of score categories per item. Tables 1 and 2 provide specifications for the TOEIC Speaking and Writing tests.

Table 1

TOEIC® Speaking Test Outline

	Item type	Item	Task	Evaluation criteria	Score
Claim 1	Read aloud	1 2	Read a text aloud	Pronunciation intonation stress	Pronunciation: 0–3 scale Intonation and stress: 0–3 scale
	Picture	3	Describe a picture	All of the above, plus grammar vocabulary cohesion	0–3 scale
Claim 2	Market survey	4 5 6	Respond to questions	All of the above, plus relevance of content completeness of content	0–3 scale, each item scored independently
	Agenda	7 8 9	Respond to questions using information provided	All of the above	0–3 scale, each item scored independently
Claim 3	Voice mail message	10	Propose a solution	All of the above	0–5 scale
	Opinion	11	Express an opinion	All of the above	0–5 scale

Table 2

TOEIC® Writing Test Outline

	Item type	Item	Task	Evaluation criteria	Score
Claim 1	Sentences	1 2 3 4 5	Write a sentence based on a picture	Grammar relevance of the sentences to the pictures	0–3 scale
Claim 2	Respond to a written message	6 7	Respond to a written request	Quality and variety of your sentences vocabulary organization	0–4 scale
Claim 3	Opinion	8	Write an opinion essay	Whether your opinion is supported with reasons and/or examples grammar vocabulary organization	0–5 scale

Scoring of the *TOEIC®* Speaking and Writing Tests

Scoring for the TOEIC Speaking and Writing tests occurs independently at the item level. After all items are scored, the final scores for each item are summed to calculate a total raw score. Then a conversion table is applied to raw scores to get scaled scores that are reported to examinees. No single rater scores the whole speaking or writing test for any individual test taker. In fact, a minimum of three different raters contribute to the total score of each test taker. Thus, the influence of any individual rater on the total test score of each test taker is minimized. Raters are allowed to rate only the speaking test or the writing test, not both. Also, at each scoring shift, each rater may rate no more than two item types. Under this practice, raters do not need to frequently switch from one item type to another and apply a different scoring rubric, which makes it easier for the raters to accurately apply the same scoring rubric across time. For additional details, please refer to Everson and Hines (2010).

All the item responses are rated by human raters through the Online Scoring Network (OSN), which has the following major advantages according to Everson and Hines (2010):

1. OSN makes random selection of responses and random assignment of responses to raters easier.
2. OSN provides instant summary statistics to scoring leaders, so raters' performances are monitored instantly.
3. OSN prevents uncertified raters or raters who failed calibration tests to access the responses to be scored.
4. OSN can track the number of questions a rater has scored from an individual test taker.
5. OSN makes it easier for raters to apply the same scoring criteria consistently.
6. Qualified raters do not need to be local to a scoring center to participate.

In order to monitor interrater consistency, some responses are rated by two raters for every item. Given the inherent judgment required for scoring CR items, equally well-trained raters may not always assign exactly the same ratings to the same responses. The TOEIC Speaking or Writing scoring team considers ratings that differ by no more than 1 raw score point as an allowable difference. When scores from two raters differ by more than one score point, they are considered discrepant, and resolution by a scoring leader (a rater with additional training and experience) is required before scores are reported.

Rater Training for the TOEIC Speaking and Writing Tests Before an Official Scoring

Educational Testing Service (ETS) devotes substantial resources to rater training and monitoring during the scoring sessions to ensure accuracy of the scoring for the TOEIC Speaking and Writing tests. Raters for the TOEIC Speaking and Writing tests are required to 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. During the initial training phase, raters learn about the format of the TOEIC Speaking or Writing tests, item types, and scoring rubrics. After training, in order to become a qualified rater for the TOEIC Speaking or Writing tests, trainees enter OSN to take a certification test. If they pass the certification test, they become qualified raters for the

TOEIC Speaking or Writing tests. Otherwise, they must undergo more training and take a different certification test at a later date. For more details, please refer to Everson and Hines (2010).

Monitor Rater Performance During Each Administration

To ensure that raters understand and apply the scoring rubric accurately and consistently, calibration is required. At each scoring session, each rater is assigned to a scoring team that scores the same question. Each team has a scoring leader, whose role is to monitor the accuracy of each rater on his/her team. Each official scoring session begins with raters completing a calibration set. This session may take place at the beginning of each day, at the beginning of the scoring for a new item type, or when raters work on the same item type longer than 4 hours. The calibration set consists of a number of test-taker responses that have been reviewed by scoring experts who agree on a preset score for each of them. Each rater scores the set of responses. If a rater's scores do not agree with the assigned scores to an acceptable level of accuracy, the rater confers with the scoring leader and then scores a different calibration set. If the rater fails on the second calibration set, the rater is dismissed from scoring that day and asked to review training materials before the next scheduled scoring session. OSN prevents raters from accessing responses until they have successfully passed calibration to demonstrate that they are scoring on track. During scoring, raters can access benchmark responses, representing prototypical examples at each score level, to review or clarify points about the rubric. All raters are required to listen to or read benchmark responses before their scoring session begins and after they have returned from a break. 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 further help raters understand the scoring criteria.

During the scoring shift, scoring leaders monitor the raters primarily through back scoring, a process by which they blindly review responses that a rater has scored and, if needed, work with the rater to remediate any error or misunderstanding of the rubrics. The assigned score is changed to the correct score. Scoring leaders are available to answer questions that raters may have (talking by phone and by a chat function that is internal to OSN for security purposes) and also to assist raters in scoring responses that are unusual or difficult to score. If, when using the various monitoring tools available to them, scoring leaders find a rater who is consistently scoring off-target during a scoring session, all of the rater's scores can be cancelled and the responses rescored.

Scoring leaders also prepare a daily end-of-day report that summarizes both OSN-related issues and content issues. The scoring leader notes any questions related to prompts, difficult-to-score responses, or rubrics that came up during the day. Raters' performance is also monitored by content scoring leaders (CSLs), who are more experienced than scoring leaders. CSLs mentor new scoring leaders on difficult-to-score responses. They also compile content information from scoring leaders' end-of-day reports and report any potential content flags to ETS Assessment Development (AD) team. This information helps AD to revise future items, write better topic notes to the scoring rubric, and provide better training.

Agreement rates for scoring consistency of each item are calculated after scoring is finished and before scores are reported based on responses rated by two raters. During each administration, items with low agreement rate are flagged for inspection. AD investigates the scoring of these items by checking the scoring accuracy of some randomly selected responses, especially responses rated by new raters. Average agreement rates (i.e., percentage of double-rated responses with allowable difference in the two ratings) for each item type in the TOEIC Speaking and Writing tests based on data from September 2012 to January 2013 are presented in Tables 3 and 4.

Table 3

Agreement Rate Based on Data From September 2012 to January 2013 for Speaking

Item	Agreement rate
1 – Into.	99.94
1 – Pro.	99.97
2 – Into.	99.96
2 – Pro.	100.00
3	99.66
4	99.75
5	99.30
6	99.62
7	99.78
8	99.90
9	100.00
10	98.42
11	99.13

Table 4

Agreement Rate Based on Data From September 2012 to January 2013 for TOEIC Writing

Item	Agreement rate
1	99.54
2	99.64
3	99.50
4	99.43
5	99.64
6	97.78
7	96.84
8	99.70

Post-Administration Procedures for Monitoring Individual Rater Performance

In addition to agreement rate, which reflects the overall rater performance at item level, the ETS Statistical Analysis (SA) team runs analyses based on responses with double ratings in a 3-month period to identify individual raters whose scoring behavior is inconsistent with that of other raters such that additional training can be provided to these individuals. Individual raters' scoring leniency/severity or scoring scale preferences are evaluated by comparing individual rater means, standard deviations, and score distributions to the final ratings of the same responses, which can be from different items in different forms.

For each rater, SA calculates the difference between his/her average ratings and the average of the final ratings, the variance ratio (VR) of his/her ratings and the final ratings, and the difference between his/her percentage of each score category and the percentage of each score category based on the final ratings. If the difference in the average ratings falls beyond the 95% confidence interval of its mean, the rater is flagged, either as MN_H (high mean score, which means the rater awards higher scores on average) or MN_L (low mean score, which means the rater assigns lower scores on average). If the VR of ratings for a rater falls beyond the 95% confidence interval of its mean, the rater is flagged as either VR_H (meaning the rater's ratings are more spread out) or VR_L (meaning the rater's ratings are more clustered together, indicating that he or she may not use all the score categories). If the difference in the percentage choosing a score of 5 for a rater is significantly higher than the mean of such differences across all raters, the rater is flagged as 5_H (the rater awards score 5 more often than other raters). On the other hand, if the difference of percent choosing score 5 is significantly lower than its mean, the rater is flagged as 5_L (the rater seldom uses category 5). If a rater has both flags VR_L and 3_H (the rater uses category 3 more often than other raters), then this rater tends to use the score categories in the middle of the scale. Individual raters are also flagged if their ratings are different from the final scores by more than 1 point, or if their exact agreement rate is significantly lower or higher than that of other raters. SA also summarizes how many times a rater's rating is discrepant from the final rating. For the item types with score categories from 0 to 5, the total number of flags can be 11; for item types with score categories from 0 to 4, the total number of flags can be 10. An example output file for flagging individual raters is provided in Table 5. Rater 1 had both VR_L and 3_H flags, suggesting that this rater tends to assign scores in the middle. Raters 2 and 3 both had discrepancies with final scores. AD and scoring leaders will monitor these raters closely in the future. Raters whose ratings are frequently discrepant from the final scores are brought to the attention of scoring leaders. This type of flag can provide accuracy information on individual rater's scoring performance since all the final scores for responses with discrepant ratings are provided by expert raters whose ratings can be considered as accurate. It is important to note that these flags are merely suggestive of the need for further investigation. It is possible, by chance assignment, that a rater may encounter a set of responses over a period of time that are deserving of lower scores than the average pool of raters, and therefore, their scoring would be accurate. Additional back reading and monitoring are necessary to determine if a rater is in fact having a problem with scoring accurately.

Table 5

Example Output for Flagging Individual Raters

Rater	Number of rated responses	Mean	STD	Percentage distribution of ratings						Exact agreement rate	Discrepancy rate	Total number of flags	Flags
				0	1	2	3	4	5				
1	39	2.90	0.38	0.00	0.00	12.82	84.62	2.56	0.00	87.18	0.00	4	VR_L, 3_H, 4_L, Corr_L
Final ratings	39	2.97	0.54	0.00	0.00	15.38	71.79	12.82	0.00				
2	52	3.23	0.78	0.00	1.92	7.69	63.46	19.23	7.69	78.85	1.92	3	MN_H, 5_H, Discrepancy
Final ratings	52	3.04	0.71	0.00	3.85	9.62	67.31	17.31	1.92				
3	46	2.85	0.97	0.00	8.70	23.91	45.65	17.39	4.35	80.43	2.17	3	MN_L, 5_L, Discrepancy
Final ratings	46	3.02	1.00	0.00	6.52	19.57	47.83	17.39	8.70				
4	102	2.79	0.81	0.00	5.88	27.45	48.04	18.63	0.00	82.35	0.00	0	
Final ratings	102	2.75	0.79	0.00	5.88	28.43	50.00	15.69	0.00				
5	202	2.82	0.77	0.00	3.47	26.24	59.41	6.93	3.96	87.62	0.00	0	
Final ratings	202	2.78	0.75	0.00	3.47	28.22	57.92	7.43	2.97				

Notes. VR_L = low variance ratio, which means a rater uses a narrow score range; 3-H = high percentage of score 3 comparing to final ratings; 4_L = low percentage of score 4 comparing to final ratings; Corr_L = lower interrater correlation; MN_H = high average score comparing to the average of the final ratings; Discrepancy = a rater's rating differs from the final rating by more than 1 point.

The SA team also provides summative information about the number of forms rated by each rater, the number of responses rated by each rater, the number of flagged forms and flagged responses for each rater, and the total number of flags. The AD staff keeps track of raters who are flagged more often or on more forms, and then attempt to determine the reason for the discrepant performance in order to provide extra training for these raters.

Future Directions for Monitoring and Enhancing Scoring Quality

Interspersing validity (or monitor) papers (papers that have been prerated by expert judges) into each administration can provide a true comparison baseline for evaluating raters' performance (Johnson, Penny, & Gordon, 2009). For forms that do not reuse any items from previous forms, scoring leaders can prerate a random sample of responses before the scoring session begins and treat these responses as monitor papers.

Summary

The TOEIC Speaking and Writing tests utilize a variety of test question types that require test takers to construct responses, not simply to choose among prespecified options. Because these responses are subjectively scored by human raters, there is a possibility that human error can reduce the accuracy of test scores. The *TOEIC*® program employs multiple carefully developed procedures to monitor rater performance in order to ensure that potential human error is kept to a minimum. Item level scoring, calibration, benchmark responses, and topic notes help raters to understand the scoring rubric accurately and to apply the same scoring criteria consistently over time. Back reading helps scoring leaders monitor raters' performance in time and improves scoring accuracy. Post-hoc rater monitoring also provides useful information about an individual rater's performance, which in turn helps with rater training and monitoring and protects score accuracy and quality.

References

Allalouf, A. (2007). Quality control procedures in the scoring, equating, and reporting of test scores. *Educational Measurement: Issues and Practice*, *26*, 36–46.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*(4), 289–303.

Everson, P., & Hines, S. (2010). How ETS scores the TOEIC Speaking and Writing test responses. *The research foundation for TOEIC: A compendium of studies* (pp. 8.1–8.9). Princeton, NJ: Educational Testing Service.

Fitzpatrick, A. R., Ercikan, K., & Yen, W. M. (1998). The consistency between raters scoring in different test years. *Applied Measurement in Education, 11*, 195–208.

Hoskens, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed-response items: An example from the golden state examination. *Journal of Educational Measurement, 38*, 121–145.

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. New York, NY: The Guilford Press.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*, 413–428.