Measuring the Power of Learning.®

Research Memorandum
ETS RM–17-04

Linking *TOEIC*® Speaking Scores Using TOEIC Listening Scores

Sooyeon Kim

December 2017

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Linking *TOEIC*® Speaking Scores Using TOEIC Listening Scores

Sooeyon Kim
Educational Testing Service, Princeton, New Jersey

December 2017

Corresponding author: S. Kim E-mail: skim@ets.org

Suggested citation: Kim, S. (2017). *Linking* TOEIC® *Speaking scores using TOEIC Listening scores* (Research Memorandum No. RM-17-04). Princeton, NJ: Educational Testing Service.

Action Editor: Donald Powers

Reviewers: Yanmei Li and Yanxuan Qu

Abstract

The purpose of this study was to assess the effectiveness of the current practice for reporting scores on the *TOEIC*® Speaking test. Currently test developers adhere to strict specifications to ensure that each new edition (or form) of the TOEIC Speaking test is comparable to previously used forms in terms of content and difficulty. For each of the 30 TOEIC Speaking test forms, the operational scores derived from the current practice were compared to the scores derived from the external multiple-choice (MC) anchor linking design. Scores on the TOEIC Listening test were used as an external anchor test for linking. Score differences derived by the two procedures were generally minimal and comparable to differences resulting from measurement error. The study suggests that psychometric benefits that may be achieved by replacing the current practice with an external MC anchor linking design will be minimal. The TOEIC program currently conducts various statistical checks against the historical records in an attempt to maintain score comparability over forms and over time. The continuation of the checking procedure will be a practical choice for this test to maintain comparability of the reporting scale over time.

Key words: constructed response, external anchor linking, test fairness, human rating

In developing multiple forms of a test, test developers use test specifications to ensure that the alternate forms are similar in content and statistical characteristics. As well specified as the test development process may be, typically, slight differences often occur in the statistical difficulty of the alternate forms. For tests containing constructed-response (CR) items that require test takers to construct responses (instead of selecting them from multiple choices), the specifications must also include a scoring rubric for each item, which must be consistently applied by the raters when the CR items are employed in different test forms or administrations. Even so, CR items bring certain complications in that rater standards may shift from one administration to another, even if the scoring rubric has not changed. Thus one form can be more difficult than another due to either (a) the inclusion of more difficult items, (b) more stringent scoring by raters, or (c) both. Under these circumstances, scores on one form would not indicate the same level of ability as the same scores on another. Test equating is a statistical method for adjusting for difference in difficulty among forms that are built to the same specifications. Various equating designs and methods have been discussed thoroughly in the literature (Kolen & Brennan, 2004). Perhaps most often, equating occurs in the context of the nonequivalent groups with anchor test (NEAT) design, in which a set of items common to both the new and reference forms is used to place both forms on the same scale.

In using a NEAT design, a major drawback with tests comprising CR items is the difficulty of identifying a satisfactory anchor test. In many cases, for example, CR items are not reused across different test forms because of ease of memorization (Muraki, Hombo, & Lee, 2000), so that there are no common CR items available for equating. Even if CR items were reused, the CR anchor items may not behave in the same way in both testing groups over time, because raters might change their scoring standards from one time to the next. Thus use of common CR items, which are not strictly equivalent, would lead to erroneous results (Kim, Walker, & McHale, 2010b; Tate, 1999). Some practitioners have suggested using MC items as anchors to adjust for differences in difficulty among test forms containing CR items (e.g., Baghi, Bent, DeLain, & Hennings, 1995; Ercikan et al., 1998). Evidence suggests, however, that using an all-MC anchor with tests made up of CR items will lead to biased equating results (Kim & Kolen, 2006; Kim, Walker, & McHale, 2010a; Li, Lissitz, & Yang, 1999), possibly because the MC and CR items may measure somewhat different constructs (Bennett, Rock, & Wang, 1991; Sykes, Hou, Hanson, & Wang, 2002). For those limitations, many testing programs carry out

routine statistical procedures (e.g., monitor raters' scoring behaviors or item difficulty) instead of equating in an attempt to maintain score comparability over forms and administrations.

The *TOEIC*® Speaking Test

The *TOEIC*® tests are English language proficiency tests for people whose native language is not English. The TOEIC Speaking test is intended to measure the test taker's ability to communicate in spoken English in daily life and in the workplace. The test consists of 11 items, representing 6 types of speaking tasks, requiring about 20 minutes to complete. The type of task and rating scale are presented in Table 1. For security reasons, all of the TOEIC Speaking test forms include newly developed items only, and thus no common CR items exist across any forms.

Table 1 Test Specifications of TOEIC Speaking Test

Item	Task	Rating scale
1–2	Read a text aloud	Intonation: 0–3; Pronunciation: 0–3
3	Describe a picture	0–3
4–6	Respond to questions	0–3
7–9	Respond to questions using information provided	0–3
10	Propose a solution	0–5
11	Express an opinion	0–5

The scaled scores of the TOEIC Speaking test range from 0 to 200 in increments of 10. The comparability of the scores across forms of the TOEIC Speaking test is mainly controlled through consistent item development and scoring. Because it is often difficult to achieve these conditions constantly in practice, however, the TOEIC family of products and services routinely exercises additional statistical checks to enhance the score comparability across forms.

Purpose

The major purpose of this study is to assess the effectiveness of the current practice of applying a single scale score conversion to all new editions of the test. To that end, a comparison of the scores resulting from a linking design and the current practice was made. Score conversions based on the NEAT design through TOEIC Listening test scores (external MC anchor) were derived from 30 operationally administered forms of the TOEIC Speaking test. The conversions resulting from a conventional linking procedure were then compared to the operational conversion resulting from the current practice to compute score differences resulting from two different procedures. In practice, the true relationship between any two TOEIC

Speaking test forms is unknown, and thus this comparison cannot lead to a definitive conclusion as to which procedure is a better choice for the TOEIC Speaking test. The magnitude of the score differences between the two procedures could be used as a gauge to assess the effectiveness of the current practice.

Method

Data

For this study, test takers' records were gathered from the TOEIC Speaking test forms that had been administered between February 2014 and November 2015. The 60 forms taken by a large number of test takers (e.g., more than 1,000) were designated as either a new form (30 forms) or a reference form (30 forms). None of these forms shared items with another, so the choice regarding which new form to link to which reference form was somewhat arbitrary. However, I attempted to mimic the real world by linking more recently administered forms to older forms. In general, the gap between the new and reference form administrations was 3–6 months.

Table 2 presents the descriptive statistics of the TOEIC Speaking test scores and the TOEIC Listening test anchor scores in both new and reference form groups at each administration. The correlations between the TOEIC Speaking test (CR) scores and Listening test (MC) anchor scores are also included in Table 2. As the anchor standardized mean differences (SMDs) indicate, the reference group was more able than the new group on 22 forms out of 30 (e.g., SMD $\leq -.1$). The size of the difference between the MC anchor means of the new and reference form groups varied from −.36 to .24 in standard deviation units. The correlations between the CR score and MC anchor scores ranged from .57 to .71 (M = .63, SD= .03). As expected, the anchor correlations are not as high as the anchor correlations usually observed in the MC-only test equating using an internal anchor (r = .80 or higher).

Table 2 Means and Standard Deviations of the TOEIC Speaking Test and Listening Test Scores in the New and Reference Form Groups

Speaking form	NF *N*	NF Speaking, *M* (*SD*)	NF anchor, *M* (*SD*)	NF *r*	RF *N*	RF Speaking, *M* (*SD*)	RF anchor, *M* (*SD*)	RF *r*	SMD (new-ref)
1	1,321	130 (20.7)	387 (69.3)	0.66	1,666	124 (21.3)	402 (68.9)	0.62	−0.22
2	1,482	126 (21.0)	393 (66.9)	0.63	2,053	129 (22.4)	396 (68.0)	0.64	−0.03
3	1,350	130 (21.0)	389 (66.1)	0.62	1,503	126 (21.2)	404 (66.5)	0.63	−0.22
4	1,695	129 (22.9)	395 (68.9)	0.62	1,651	126 (22.8)	404 (69.7)	0.66	−0.13
5	1,868	125 (21.3)	395 (67.0)	0.61	1,731	124(22.6)	411 (64.0)	0.62	−0.24
6	1,337	129 (20.0)	396 (68.8)	0.60	1,719	132 (20.2)	408 (66.8)	0.61	−0.18
7	1,569	127 (21.0)	396 (69.7)	0.63	1,546	121 (22.6)	404 (70.4)	0.62	−0.10
8	1,959	128 (22.1)	396 (67.8)	0.65	1,540	130 (21.4)	406 (68.9)	0.61	−0.15
9	1,986	128 (20.0)	398 (68.0)	0.60	1,427	119 (21.9)	396 (72.7)	0.63	0.03
10	1,596	120 (21.3)	394 (69.5)	0.63	1,334	129 (22.1)	409 (67.7)	0.64	−0.22
11	1,588	129 (20.7)	395 (67.5)	0.62	1,814	123 (21.5)	406 (68.8)	0.63	−0.17
12	1,353	124 (22.4)	396 (68.2)	0.67	1,384	127 (21.4)	409 (66.3)	0.65	−0.19
13	1,938	126 (20.8)	393 (70.4)	0.63	1,306	125 (20.8)	408 (65.7)	0.60	−0.22
14	2,057	126 (21.3)	395 (66.8)	0.62	1,328	121 (23.4)	405 (69.2)	0.66	−0.14
15	1,939	128 (21.1)	391 (70.0)	0.66	1,472	127 (22.8)	398 (70.9)	0.63	−0.11
16	1,671	128 (19.7)	396 (68.7)	0.60	1,345	121 (21.3)	402 (67.1)	0.57	−0.09
17	1,805	127 (21.3)	396 (68.0)	0.60	1,419	127 (22.0)	397 (70.6)	0.62	−0.01
18	1,203	124 (20.6)	386 (69.9)	0.58	1,674	121 (21.9)	396 (70.4)	0.65	−0.13
19	1,371	125 (21.4)	385 (71.0)	0.62	1,202	130 (21.0)	410 (65.5)	0.65	−0.36
20	1,362	124 (22.5)	388 (72.7)	0.62	1,205	121 (23.6)	396 (75.1)	0.64	−0.10
21	1,203	130 (20.9)	408 (65.8)	0.57	1,325	128 (22.3)	401 (74.6)	0.66	0.09
22	1,139	127 (19.8)	387 (67.4)	0.61	1,223	130 (21.9)	405 (68.0)	0.62	−0.27
23	1,276	127 (21.3)	398 (67.0)	0.64	1,213	125 (20.4)	397 (67.9)	0.62	0.02
24	1,307	121 (20.7)	394 (71.0)	0.63	1,320	135 (18.7)	411 (64.7)	0.56	−0.26
25	1,419	129 (21.6)	392 (68.8)	0.64	1,178	117 (24.5)	393 (72.8)	0.66	−0.01
26	1,297	126 (20.2)	394 (67.7)	0.59	1,286	127 (22.9)	401 (70.8)	0.62	−0.11
27	1,332	121 (22.2)	392 (69.7)	0.63	1,430	127 (21.3)	404 (66.1)	0.59	−0.18
28	1,256	124 (21.8)	390 (73.2)	0.63	1,412	125 (21.7)	398 (69.2)	0.64	−0.10
29	1,331	130 (21.7)	398 (71.8)	0.63	1,327	128 (22.6)	407 (68.6)	0.59	−0.13
30	1,498	127 (22.2)	403 (69.8)	0.63	1,282	125 (25.0)	386 (75.6)	0.71	0.24

Note. r = correlation between TOEIC Speaking and Listening test (anchor) scores. Not all test takers' anchor scores were available at the time of linking. NF = new form linking group, RF = reference form linking group, SMD = standardized mean difference between the MC anchor scores (new group minus reference group).

Procedure

For each form, score linking through the TOEIC Listening test scores was conducted using the test takers whose TOEIC Listening test scores were available at the time of linking.[1] On average, TOEIC Listening test scores were available for approximately two-thirds of test takers (M = 67%, range = 50%–78%).[2] In the NEAT design with external MC anchor, the chained equipercentile (Kolen & Brennan, 2004) method was used to produce the scaled sore conversion.[3] The resulting conversion was then applied to every test taker in the new form group to obtain his or her TOEIC Speaking test scaled score. This scaled score is the scaled score the test taker would have received if the linking design with an external MC anchor had been implemented in the operational setting. I computed the difference between the new scaled score based on the MC linking design and the test taker's operational scaled score based on the current practice. Then I computed the percentage of test takers whose scaled scores were categorized as follows: no difference, a 10-point difference, a 20-point difference, and so on. In addition, the means and standard deviations from all test takers in the new form group were calculated based on the two sets of scaled scores, along with the SMDs (linking minus operational in the new group).

Results

Table 3 presents the scaled score difference (external linking conversion minus operational conversion) results associated with the 30 TOEIC Speaking test forms investigated in this study. As shown, the results are highly consistent across all the 30 forms, indicating similar differences. The differences were primarily within the range of −10 to +10. For 14 of the forms (Forms 2, 6, 8, 12, 15, 17, 19, 21–23, 26–28, and 30), the scaled scores remained unchanged for more than two-thirds of the test takers. On 10 forms (Forms 1, 3, 5, 7, 9, 11, 14, 16, 18, and 25), more than 85% of the test takers' scaled scores decreased by 10 points when the linking conversion was applied. The linking conversion, however, led to a 10-point increment for 95% of the test takers who took Form 24. The differences on the remaining forms were rather evenly distributed across two adjacent categories, either −10 to 0 (Forms 4, 13, 20, and 29) or 0 to +10 (Form 10). For those five forms, approximately 33%–50% of the test takers retained the same scaled scores, whereas approximately 50%–63% of the test takers' scaled scores changed by 10 points. Very few test takers' score differences were greater than 10 points.

Table 3 Differences Between External Anchor Linking Scaled Scores and Operational Scaled Scores for the 30 Test Forms

Form	Below −20	−20	−10	0	10	20	Above 20
1	–	0.7	96.7	2.2	0.4	–	–
2	–	–	–	98.8	1.2	–	–
3	–	0.1	86.2	13.7	–	–	–
4	–	–	49.2	50.8	–	–	–
5	–	–	97.5	2.5	–	–	–
6	–	–	0.1	99.9	–	–	–
7	–	0.1	85.1	14.8	–	–	–
8	–	–	–	99.1	0.9	–	–
9	0.2	–	99.6	0.3	–	–	–
10	–	–	–	41.4	58.6	–	–
11	0.3	–	94.0	5.7	–	–	–
12	–	0.4	7.0	90.8	1.7	0.1	–
13	–	–	45.9	53.2	0.7	0.1	0.1
14	–	–	98.8	1.1	0.2	–	–
15	0.1	–	29.0	70.9	–	–	–
16	0.5	–	99.1	0.4	–	–	–
17	–	0.3	2.0	97.7	–	–	–
18	–	–	89.6	10.3	0.1	–	–
19	–	–	0.7	99.3	–	–	–
20	–	–	56.2	43.7	0.1	–	–
21	0.3	0.2	0.4	99.2	–	–	–
22	–	–	–	99.1	0.8	0.1	–
23	–	–	–	99.7	0.3	0.1	–
24	–	–	–	3.8	95.5	0.6	0.2
25	–	–	99.4	0.6	–	–	–
26	–	–	–	95.3	4.7	0.1	–
27	–	–	–	98.8	1.2	–	–
28	–	–	7.4	92.6	–	–	–
29	–	–	62.7	33.0	4.1	0.2	–
30	–	–	–	100.0	–	–	–

Table 4 presents the means and standard deviations of the operational and linked scaled scores computed from all test takers in the new form group. The total group also includes the test takers who were excluded from the linking process because their anchor scores were not available at the time of linking. The direction of mean differences between the two conversions was matched with the difference direction shown in Table 3. As expected, the SMDs of the score differences between the current practice and MC anchor linking was close to half of a standard

deviation (around 10 points) under the nine form cases (Forms 1, 5, 9, 11, 14, 16, 18, 24, and 25). For many forms used in this study, the MC–anchor linking produced lower means than did the current practice. The largest mean difference between the two conditions was 10 points, which is slightly lower than one-half of a standard deviation.

Table 4 Means and Standard Deviations of the Scaled Scores Derived From Chained Equipercentile Linking and Current Operational Practice

Speaking form	*N*	Current practice *M*	Current practice *SD*	Chained equipercentile *M*	Chained equipercentile *SD*	SMD
1	2,621	130	22.9	120	22.0	−0.44
2	2,419	125	22.1	125	22.4	0.00
3	2,310	130	23.3	121	21.3	−0.39
4	2,629	128	24.0	123	22.8	−0.21
5	2,887	125	22.3	115	21.8	−0.44
6	1,885	128	21.2	128	21.2	0.00
7	2,258	126	21.7	118	22.4	−0.39
8	2,794	128	22.9	128	22.7	0.00
9	2,859	127	20.9	117	21.4	−0.47
10	2,394	119	22.5	125	20.9	0.27
11	2,403	129	22.0	119	21.5	−0.44
12	2,031	123	23.7	122	21.8	−0.03
13	2,641	125	21.9	121	23.5	−0.20
14	2,730	125	22.1	115	22.0	−0.45
15	2,717	128	22.3	125	24.6	−0.13
16	2,226	127	20.9	117	21.5	−0.48
17	2,432	127	22.3	126	22.2	−0.01
18	1,660	123	21.9	114	22.7	−0.40
19	1,916	125	22.6	125	22.4	0.00
20	1,902	124	23.2	118	21.0	−0.25
21	1,545	128	22.6	128	23.3	−0.01
22	2,312	127	22.2	127	21.9	0.00
23	2,016	125	22.3	125	22.2	0.00
24	1,708	120	21.8	130	20.8	0.46
25	2,466	128	23.7	118	23.9	−0.42
26	1,982	126	22.0	126	22.6	0.02
27	1,973	122	23.1	122	22.8	0.01
28	1,990	122	23.9	121	24.5	−0.03
29	2,047	128	23.8	122	23.9	−0.24
30	2,013	126	23.2	126	23.2	0.00

Note. SMD = standardized mean difference.

To display the score region where most test takers were located, the scaled score distributions of the new form group accumulated over the 30 forms/administrations are presented in Figure 1. In the figure, one distribution was associated with the relative frequency from the current practice, and another was associated with the relative frequency from the MC–anchor linking. Figure 2 plots the differences from the operational conversion across the scaled score region from the 1st percentile to the 99th percentile in the new form group. There were 30 difference lines associated with the 30 forms, and the dotted lines at ±10 indicate half of a standard deviation. The differences were generally smaller than 10 points across the score region where most test takers were located. However, the difference line associated with Form 25 (solid blue line) was beyond the ±10 band. In Form 25, the new form group was as able as the reference group, as the SMD of the TOEIC Listening test score indicates (SMD = −.01). However, the TOEIC Speaking test score of the new form group (M = 129) was much higher than that of the reference form group (M = 117), leading to the SMD of .53. Because the TOEIC Speaking test form difference in difficulty was adjusted through the TOEIC Listening test scores under the MC–anchor linking design and the reference form group did as well on the TOEIC Listening test, the new TOEIC Speaking test form appeared much easier in the process of score linking.

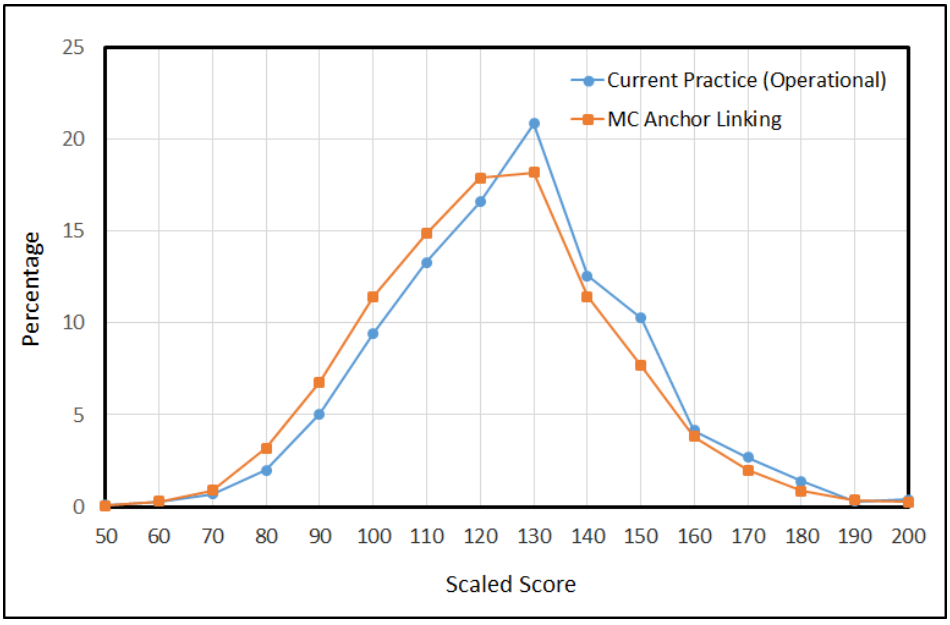

Figure 1. Percentage distribution of the scaled scores on the entire new form group.

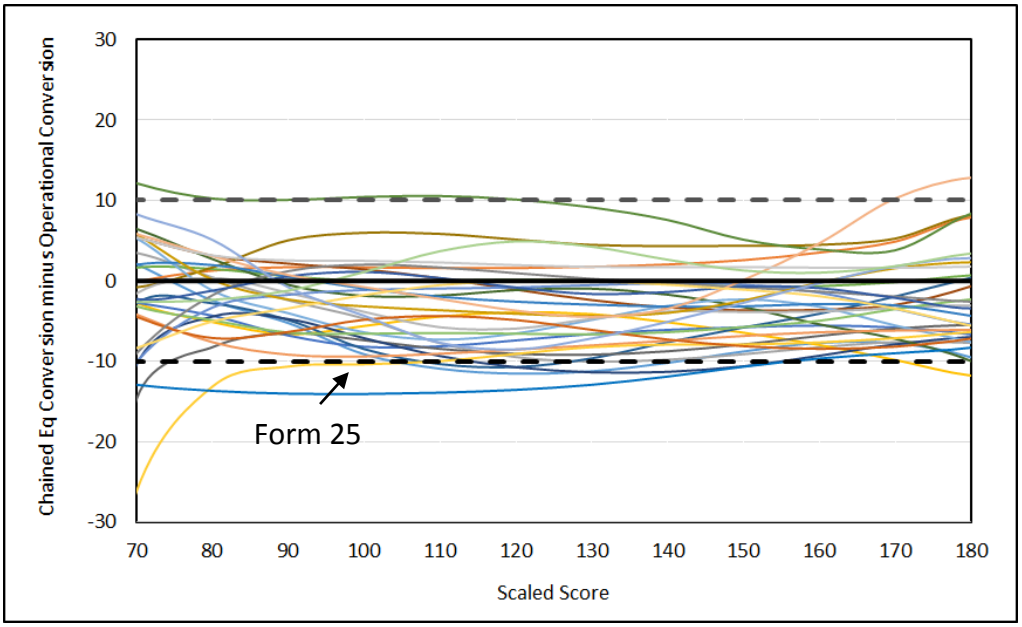

Figure 2. Difference plots between chained equipercentile conversion and operational conversion.

Discussion

Owing to security concerns, the TOEIC family of products and service uses new editions of the TOEIC Speaking test (which include only newly developed CR items) for every test administration. To ensure score comparability over different forms, some form of equating is desirable. When test forms consist of CR items only, however, score equating through the conventional design (e.g., NEAT) is not always feasible because of a lack of proper common items. Although the TOEIC Listening and Reading test can be used as an external MC anchor to link the TOEIC Speaking test scores, using an external MC anchor can potentially be problematic in that anchors consisting of external MC items alone may not adequately represent the CR test content and thus may not produce satisfactory links. In addition, because not all test takers' external MC anchor scores are available at the time of linking, it is often questionable how well a linking sample represents the entire group of test takers. Owing to various practical limitations (e.g., no common CR items, low volume, operational demands for reporting scores in a short time), the current practice of the TOEIC Speaking test is based on the assumptions that forms are sufficiently similar in difficulty and that raters use the same scoring standard, as is intended.

The purpose of the study was to compare the current practice to a procedure by which scores are derived from an MC external anchor linking design. The external MC anchor linking may not be optimal unless the correlation between MC and CR is substantially high. Even so, some testing programs use this approach operationally as a method to produce comparable CR scores over the forms. Because the external anchor scores were available for many of the TOEIC Speaking test takers and the correlations between MC and CR were moderate, the conversions derived from the external MC anchor linking were used to assess the effectiveness of the current practice in this study.

The most interesting feature of the study's results was the comparability of scores derived from different procedures. The scores derived through the external MC anchor design were generally comparable to the current reported scores based on the consistency of both item difficulty and scoring. Although some test takers' scaled scores changed by as much as 10 points, this change is comparable to measurement error, as the standard error of measurement of the TOEIC Speaking test is approximately 10–11. The present findings indicate that adopting external MC anchor design in the operational setting would have little practical impact and may therefore be unnecessary. This study suggests that the psychometric benefits that may be achieved by replacing the current practice with an external MC anchor linking may be negligible. Given the moderate correlation between TOEIC Speaking test and Listening test scores, the linkage between the two sets of scores will be weak, thus yielding minimal benefit to improved equivalence across forms.

The TOEIC family of products and service uses several strategies in an attempt to maintain score comparability over different forms and administrations. Test developers exercise their expertise to assemble the TOEIC Speaking test forms to be as parallel as possible. Because difficulty levels of CR items are determined as a function of item–rater combinations, however, CR scoring, either stringent or lenient, may change the level of item difficulty as well. Often slight scoring shifts over time are unavoidable. In the TOEIC family of products and service, all raters are thoroughly trained in the use of the rating rubrics to enhance the consistency of the ratings and, therefore, the reliability of the TOEIC Speaking test scores. Raters are required to pass a certification test, consisting of a number of benchmark responses for which consensus ratings exist, prior to starting the rating of operational responses. Expert raters provide ongoing monitoring of their ratings and are also available to provide support and feedback as needed.

The TOEIC family of products and service conducts comprehensive postadministration analyses on every administration to (a) evaluate the quality of the ratings, (b) assess the statistical/psychometric properties of each item, and (c) monitor test takers' performance over time. For example, rating consistency per item, indicated by the correlation between the two ratings and the weighted kappa coefficient (Fleiss, Cohen, & Everitt, 1969), is calculated based on the 10%–15% of double scoring data. Using the double scoring data, rating agreement per item, indicated by the percentage of same rating (no difference), the percentage of adjacent ratings (1 point difference), and the percentage of discrepant ratings (more than 1 point difference), is also examined to ensure raters' scoring consistency. Such analyses are particularly helpful in evaluating the need for additional rater training. Furthermore, descriptive statistics of each item and psychometric properties of each item type are assessed against the historical data accumulated over several years. Empirical evidence, such as historical charts and data, help to inform judgment regarding the current forms' performance. Such monitoring and analyses provide relevant empirical evidence for scoring stability and test fairness.

References

Baghi, H., Bent, P., DeLain, M., & Hennings, S. (1995, April). *A comparison of the results from two equatings for performance-based student assessments*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement, 28*, 77–92. https://doi.org/10.1111/j.1745-3984.1991.tb00345.x

Ercikan, K., Schwarz, R., Julian, M. W., Burket, G. R., Weber, M. W., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response test item type. *Journal of Educational Measurement, 35*, 137–154. https://doi.org/10.1111/j.1745-3984.1998.tb00531.x

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72,* 323–327. https://doi.org/10.1037/h0028106

Kim, S., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education, 19,* 357–381. https://doi.org/10.1207/s15324818ame1904_7

Kim, S., Walker, M. E., & McHale, F. (2010a). Comparisons among designs for equating mixed-format tests in large scale assessments. *Journal of Educational Measurement, 47,* 36–53. https://doi.org/10.1111/j.1745-3984.2009.00098.x

Kim, S., Walker, M. E., & McHale, F. (2010b). Investigation the effectiveness of equating designs for constructed response tests in large scale assessment. *Journal of Educational Measurement, 47,* 186–201. https://doi.org/10.1111/j.1745-3984.2010.00108.x

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer. https://doi.org/10.1007/978-1-4757-4310-4

Li, Y. H., Lissitz, R. W., & Yang, Y. N. (1999). *Estimating IRT equating coefficients for tests with polytomously and dichotomously scored items*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Muraki, E., Hombo, C. M., & Lee, Y.-W. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement, 24*, 325–337. https://doi.org/10.1177/01466210022031787

Sykes, R. C., Hou, L., Hanson, B., & Wang, Z. (2002). *Multidimensionality and the equating of a mixed-format math examination*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Tate, R. L. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal of Educational Measurement, 36,* 336–346. https://doi.org/10.1111/j.1745-3984.1999.tb00560.x

Notes

[1] The TOEIC Listening and Reading Comprehension (LC & RC) MC test scores can be used as an external MC anchor to link the Speaking score. Because the LC section scores showed slightly higher correlations with the Speaking scores than did either RC or LC & RC combined, the LC section score was used as an anchor in the study. However, the same trend appeared in both (RC only and LC & RC combined) anchor conditions.

[2] There exists substantial overlap between the TOEIC Speaking test population and the TOEIC LC & RC test population. However, not all TOEIC Speaking test test takers take the TOEIC LC & RC test, and vice versa.

[3] Frequency estimation equipercentile (often called poststratification equipercentile [PSE]; Kolen & Brennan, 2004, pp. 135–143) was also used to produce the conversion table for each of the 30 forms. Because both chained equipercentile and PSE produced very similar results, the PSE results were not presented in this report for simplicity. The frequency estimation equipercentile results are available on request.