Measuring the Power of Learning.®

Research Report
ETS RR–17-46

The Consistency of *TOEIC*® Speaking Scores Across Ratings and Tasks

Jonathan E. Schmidgall

December 2017

ETS Research Report 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 Report 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 Report 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.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

The Consistency of *TOEIC*® Speaking Scores Across Ratings and Tasks

Jonathan E. Schmidgall

Educational Testing Service, Princeton, NJ

This report briefly reviews the design and scoring procedure for the *TOEIC*® Speaking test and summarizes existing evidence about the consistency of TOEIC Speaking test scores. It then describes several analyses conducted using generalizability theory to provide additional information about the consistency of scores across different aspects of the scoring procedure. The results of these analyses provide more robust information about consistency of TOEIC Speaking scores with respect to important facets of the assessment procedure, such as tasks and ratings. Specifically, the results provide evidence to support claims about the consistency of ratings across different levels of the scoring procedure: tasks, claims, and scale scores.

Keywords Generalizability theory; score consistency; reliability; validation argument; ratings; *TOEIC*® Speaking test

doi:10.1002/ets2.12178

An important quality of test scores is their reliability or consistency across different aspects of the measurement procedure (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). Researchers have observed that reliability is a prerequisite to validity (Haertel, 2006), and this "conventional wisdom" is made explicit in argument-based approaches to validity in which claims about score consistency underlie subsequent inferences about the interpretation of scores (e.g., Bachman & Palmer, 2010). In a validity argument, an overall claim that scores are consistent is dependent on a series of more detailed statements about specific aspects of consistency (e.g., agreement across raters). These statements (or assertions) are backed by evidence, often in the form of reliability coefficients. A test administrator's claim that test scores are consistent is supported or weakened by the extent to which the evidence supports the various assertions regarding score consistency.

An assessment use argument (Bachman & Palmer, 2010) is an argument-based approach to validity in which claims about the meaning and use of scores rest on the foundational claim that scores should be consistent. We have utilized this approach to specify claims about the measurement quality and use of *TOEIC*® test scores. In this report, I focus on the overall claim that TOEIC Speaking test scores are consistent. Table 1 summarizes the various assertions used to advance the claim that TOEIC Speaking test scores are consistent as well as the evidence that supports each assertion.

The TOEIC Speaking test requires test takers to demonstrate their English speaking ability across 11 speaking tasks that are scored by trained raters. This scoring process transforms a test taker's speaking performance into a scale score that is an indicator of his or her English speaking ability in the context of the workplace and everyday life. As stated in Table 1, the essential claim made about this scale score is that it is consistent (or reliable) across different aspects of the measurement procedure. This claim is supported by a series of assertions that are backed by evidence from the test design process and research. For example, one way to help ensure the consistency of scores is to follow administration and scoring procedures consistently. All TOEIC Speaking tests are administered on a computer that requires the use of headphones with a microphone, and test tasks are administered using a standardized format across occasions for all groups of test takers (Hines, 2010). A standardized procedure clearly specifies the steps involved in obtaining a scale score for each test taker and is carefully implemented and monitored to ensure compliance (Everson & Hines, 2010; Hines, 2010; Qu & Ricker-Pedley, 2013).

Raters themselves can be a source of either systematic bias or random error, and careful selection and training of qualified raters are critical (Brown, 2012; Engelhard, 2002). Everson and Hines (2010) described the path to becoming a TOEIC Speaking rater, which includes a number of steps designed to ensure consistent and high-quality ratings. Potential raters

Corresponding author: J. E. Schmidgall, E-mail: jschmidgall@ets.org

Table 1 Underlying Assertions and Evidence to Support the Overall Claim That TOEIC Speaking Test Scores Are Consistent

Underlying claim or assertion	Published source of evidence
Administration procedures are followed consistently.	Hines (2010)
Scoring procedures are followed consistently.	Everson and Hines (2010), Hines (2010), Qu and Ricker-Pedley (2013)
Raters are trained, certified, calibrated, and monitored.	Everson and Hines (2010)
Scores are internally consistent.	Reasonably high internal consistency (Liao & Wei, 2010)
Scores from different raters (ratings) are consistent.	Reasonably high rater agreement rates (Liao & Wei, 2010; Qu & Ricker-Pedley, 2013); reasonably high generalizability of task scores (Liao & Wei, 2010)
Scores from different test forms and occasions of testing are consistent.	Liao and Qu (2010)

must (a) be qualified professionals (college graduates with experience teaching English as a second language/English as a foreign language); (b) complete a training course (which includes reviewing the purpose of each task type, sample and benchmark responses, and written explanations of scores for responses); and then (c) pass a certification test to demonstrate their rating proficiency. Applicants who pass the certification test qualify to work as raters but must subsequently pass a calibration test prior to every scoring session. The function of the calibration test is to ensure that raters maintain consistent standards for each new scoring session. Finally, each rater's performance is monitored by a scoring leader during the scoring session. All these policies and procedures are designed to promote the consistency and accuracy of rater scoring.

The use of highly trained raters and monitoring procedures helps to reduce the random error and bias introduced by human raters, but it is still essential to empirically quantify various aspects of reliability or score consistency (AERA et al., 2014). Score consistency can be quantified in a variety of ways. Prior research has found that TOEIC Speaking test scores are internally consistent (Liao & Wei, 2010), that scores from different raters are consistent (Liao & Wei, 2010; Qu & Ricker-Pedley, 2013), and that scores from different test forms and occasions of testing are consistent (Liao & Qu, 2010).

In an analysis of TOEIC Speaking pilot test data, Liao and Wei (2010) examined the interrater reliability and internal consistency of two test forms. Interrater reliability was evaluated by looking at rater agreement for each task and by using generalizability theory (G-theory) to estimate a generalizability coefficient for each task. Internal consistency for claim scores and weighted total scores was estimated using Cronbach's alpha (Cronbach, 1951) and stratified alpha (Rajaratnam, Cronbach, & Gleser, 1965), respectively. The analysis of rater agreement found acceptably high levels of rater agreement across most tasks, with exact agreement ranging from 50% to 81% and agreement within one score point ranging from 98% to 100%. In other words, very few test takers were given scores that were more than one score point apart. Generalizability coefficients for individual tasks ranged from .58 to .91 and were reasonably high for most. Estimates of internal consistency for claim scores were slightly lower for Claim 1 (.66 – .68) and slightly higher for Claim 2 (.66 – .80) and Claim 3 (.71 – .74). The internal consistency of total scores ranged from .82 to .86, acceptable estimates according to traditional rules of thumb (Knapp & Mueller, 2010). Ultimately, because total scores are used to make interpretations about speaking ability, these estimates are the most critical.

Test takers complete a particular form of the TOEIC Speaking test on a particular occasion, but their scores should not be unduly influenced by the particular test form or occasion of testing. Liao and Qu (2010) examined the so-called alternate form test – retest reliability of TOEIC Speaking raw and scale scores across different occasions (e.g., 1 – 30 days, 31 – 60 days) and test forms. The test – retest reliability coefficients estimated across occasions of five different lengths ranged from .75 to .83, which supports the claim that scale scores are consistent across test forms and occasions.

To help stakeholders better understand the measurement facets of a TOEIC Speaking scale score, Figure 1 illustrates the design of the test.

Figure 1 should be viewed from the bottom to the top to understand how intended claims about a test taker's speaking ability informed the design of the test and the scale score that reflects these claims. The TOEIC Speaking test is designed to provide an interpretation about English speaking ability with respect to three claims: generating speech that is intelligible (Claim 1), appropriate for routine social and occupational interactions (Claim 2), and connected and sustained for typical

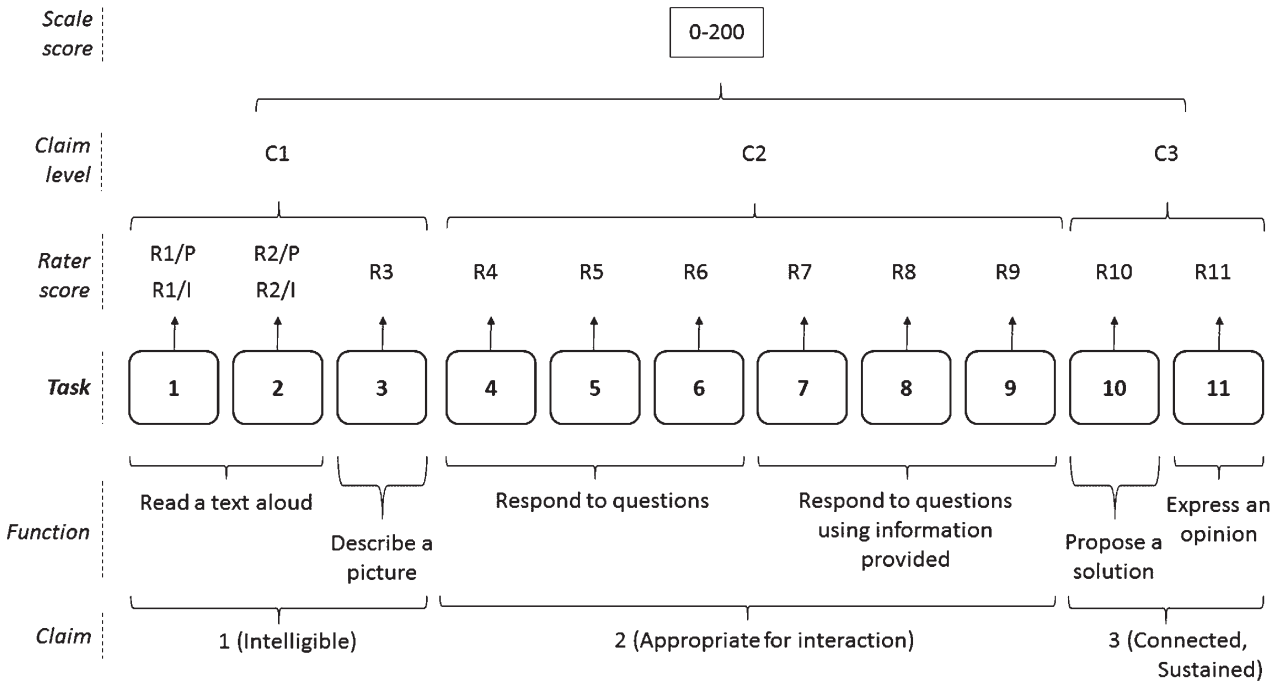

Figure 1 Design of the TOEIC Speaking test. *Note.* R = rating; C = claim; P = pronunciation subscale; I = intonation subscale.

workplace tasks (Claim 3; see Hines, 2010). As shown in Figure 1, 11 tasks were designed that targeted communicative functions that were representative of these claims. Each task is scored by a rater and assigned a single score, except for Tasks 1 and 2, which are given two scores: one for pronunciation and one for intonation. Different raters score each of the tasks, and a minimum of three different raters contribute to the final score of an individual test taker. As the figure indicates, variation in scores at the claim level reflects a test taker's performance on tasks that correspond to that claim as evaluated by different raters, for example, ratings from Tasks 1 to 3 reflect performance with respect to Claim 1. Finally, the scale score reflects a test taker's performance with respect to all three claims about speaking ability, which includes ratings of individual tasks that correspond to the claims. Ultimately, it is the scale score that is the basis for making an interpretation about someone's speaking proficiency, and so evidence of reliability or consistency is most critical for scale scores.

Research Questions

Prior research has produced evidence for the consistency of TOEIC Speaking rater scores, claim-level information, and total scores (raw or scale). However, some of that evidence is based on an analysis of pilot study data (i.e., Liao & Wei, 2010), not on operational test scores that "count." To provide updated estimates of the consistency of TOEIC Speaking scores across different phases of the scoring procedure, this study addresses the following research questions:

1. How consistent are ratings on individual tasks, as measured by generalizability and dependability coefficients?
2. How consistent is performance at the claim level across ratings, as measured by generalizability and dependability coefficients?
3. How consistent are scale scores across ratings, as measured by generalizability and dependability coefficients?

Methodology

Participants, Instrument, and Procedure

A previously administered and scored TOEIC Speaking form was rescored in its entirety. The form and set of responses that were selected to be rescored were representative of TOEIC Speaking test form administrations in terms of sample size ($N = 1{,}390$), internal consistency reliability ($\alpha = .85$), and scale score distribution ($M = 15.71$, $SD = 3.58$). Operational scoring conditions were maintained for the rescoring study (see Everson & Hines, 2010, for a description of the scoring

procedure), and raters were not aware that scoring was being performed for a research study. The number of raters scoring each set of test-taker responses varied as per operational practice but was roughly comparable across the original and rescored samples.

Analysis

The framework of G-theory (Brennan, 2001) was used to identify sources of variances associated with test-taker ability (p) and facets of the measurement procedure, which may include ratings (r') and tasks (t). The ratings and tasks facets are considered random, as they are conceptualized as representative of the population from which they are drawn without exhaustively defining it.

Although most facets of measurement are self-explanatory, a brief overview of the ratings (r') facet is needed. Each task for each person is scored by two raters, but the combination of raters differs across tasks. This approach of assigning multiple raters to each person is by design to minimize systematic bias that may arise from having the same rater or pair of raters score all of a person's responses. Thus this is a partially nested rating design in which each person is scored by multiple raters. Implementing this design using G-theory requires very large sample sizes depending on the number of rater combinations. This approach was impractical for this data set, where a large number of rater combinations was possible.

To provide a simplified approach to partially nested designs involving raters, researchers have proposed using a fully crossed design, $p \times t \times r'$, where r' represents *ratings*, not *raters* (Lee, 2006; Lee & Kantor, 2005). With ratings as a facet, some researchers have argued that a main effect cannot be interpreted as differences between people who score (raters) but simply as differences between a first and second rating (Lee, 2006). However, researchers have shown that under certain conditions, this conceptual distinction may be negligible (Lin, 2013; Sawaki, 2017; Schmidgall, 2017). For example, Lin (2013) conducted a series of simulation studies under conditions that varied sample size, number of raters, and rating conditions; he concluded that when raters are relatively homogenous (i.e., when they have similar levels of experience), the rating method is sufficient for operational use, as it sacrifices little precision. In a related effort, Schmidgall (2017) examined a number of fully crossed pairs of raters within a larger data set and found negligible rater effects, which he used to partially justify a rating method. Sawaki (2017) performed multiple analyses in which a fully crossed rating method ($p \times t \times r'$) was used for an entire data set and separate analyses were conducted for each rater pair in the data set using the rater method ($p \times t \times r$); results were largely consistent across the analyses. The purpose of the present analysis was to estimate the amount of variation across ratings irrespective of which specific raters made these ratings, so the use of this fully crossed design ($p \times t \times r'$) is appropriate.

G-theory requires the researcher to specify the relationship between the object of measurement and facets. The following sections specify the G-study designs used to estimate generalizability coefficients and variance components associated with facets of measurement for scale scores, raw scores, claim scores, individual task scores, and the five different rubrics used for the TOEIC Speaking test. G-studies were then performed using Edu-G software (Cardinet, Johnson, & Pini, 2010). Decision studies (D-studies; Brennan, 2001) were also performed to provide an estimate of reliability based on the operational scoring design of the TOEIC Speaking test, which typically uses a single rater to score each task. G-studies provide estimates that reflect the actual measurement design of a data set (e.g., 11 tasks and 2 raters), whereas D-studies provide estimates for different variations of the original measurement design (e.g., 11 tasks and 1 rater).

Individual Task Scores

Tasks 3 through 11 were assigned two ratings ($r' = 2$) using a holistic (i.e., one score) rubric, which is characterized by the design $p \times r'$. Tasks 1 and 2 use an analytic rubric in which test takers were assigned two ratings ($r' = 2$). In the scoring procedure, this results in four scores that equally contribute to the Claim 1 score. The measurement design of each of these four ratings is also $p \times r'$.

Claim-Level Performance

Performance at the claim level can be characterized by the G-study $p \times t \times r'$, in which a set of tasks are assigned two ratings ($r' = 2$). There are six tasks with Claim 2 ($t = 6$, Tasks 4–9) and two tasks with Claim 3 ($t = 2$, Tasks 10 and 11). There are three tasks associated with Claim 1 (Tasks 1–3), but four scores are produced for Tasks 1 and 2, because these

tasks are scored separately for pronunciation and intonation. Thus, for the purpose of the analyses, there are five rated tasks associated with Claim 1 ($t = 5$). This approach may introduce the halo effect and underestimate variance components associated with tasks for Claim 1, a potential limitation of this analysis.

Scale Scores

Scale scores are based on linear combinations of raw scores and can be characterized using the fully crossed G-study design $p \times r'$. Because the research question examining scale scores is concerned with consistency across occasions of ratings, task was not specified as a facet of the G-study design.

Results

Consistency of Individual Task Scores

The percentage of total variance accounted for by each facet of measurement for each task or score is summarized in Table 2. The generalizability coefficient ($\hat{\rho}^2$) based on the G-study indicates the reliability of each individual task or score assigned two ratings (i.e., for this study), whereas the coefficient based on the D-study extrapolates the reliability estimate to operational scoring conditions in which one rater is typically used to score each task.

As shown in Table 2, a person's ability (p) explains more of the variance in scores than the rating he or she received (r') or the combination of unexplained error (e) and the particular rating for a particular person (pr'). For 3 of 11 scores (1/I, 2/P, 2/I), a greater percentage of variance in scores was accounted for by the combination of unexplained error and the particular rating provided for a particular person.

Generalizability coefficients ($\hat{\rho}^2$) based on G-studies were adequate (median = .75, range, .57 – .92), particularly for Tasks 4 through 11. Generalizability coefficients based on D-studies that reflect the operational rating design ($r' = 1$) were slightly lower (median = .61; range .40 – .85). One possible explanation for the higher proportion of variance explained by the combination of error and the particular rating provided for a particular person for Tasks 1 through 3 — and, thus, lower generalizability coefficients — is restriction of range. The variance of ratings for Tasks 1 through 3 was comparatively lower than for other tasks, which may help explain the comparatively lower generalizability coefficients.

Consistency of Claim-Level Performance

The percentage of total variance accounted for by each facet of measurement for each claim is summarized in Table 3, along with generalizability coefficients based on G- and D-studies.

As seen in Table 3, most of the variance in claim-level performance was explained by ability (p); the interaction between ability and task (pt); and the combination of unexplained error (e) and the three-way interaction between person, task, and rating (ptr'). The interaction between task and ability (pt) should be interpreted as the extent to which different test takers (p) performed differently on tasks associated with that claim (e.g., Tasks 4 – 9 for Claim 2); in other words, the rank ordering of persons varied across tasks within a claim. For Claim 2, a relatively large percentage of total variance (34.4%) was explained by person – task interaction or by differences in the rank ordering of performances across different tasks. Overall, though, differences in task difficulty did not account for a high percentage of total variance (6.1%).

Table 2 Individual Task G-Study Percentage of Total Variance for Each Facet of Measurement, G-Study Generalizability Coefficient, and D-Study Generalizability Coefficient for Design With $r' = 1$

	Task/score												
Source	1/P	1/I	2/P	2/I	3	4	5	6	7	8	9	10	11
Person (p)	54.2	42.1	40.0	39.3	52.2	74.5	56.7	59.8	84.9	76.0	64.1	71.9	72.4
Rating (r')	.1	.6	.1	.5	.7	1.2	.2	1.2	.1	.3	0	0	0
pr', e	45.7	57.2	59.9	60.2	47.1	24.3	43.1	39.1	15.0	23.6	35.8	28.1	27.5
$\hat{\rho}^2$ (G-study)	.70	.60	.57	.57	.69	.86	.72	.75	.92	.87	.78	.84	.84
$\hat{\rho}^2$ (D-study)	.54	.42	.40	.40	.53	.75	.57	.61	.85	.76	.64	.72	.72

Note. P = pronunciation subscale; I = intonation subscale; G-study = generalizability study; D-study = decision study.

Table 3 Claim-Level G-Study Percentage of Total Variance for Each Facet of Measurement, G-Study Generalizability Coefficient, and D-Study Generalizability Coefficient for Design With $r' = 1$

	Claim		
Source	1	2	3
Person (p)	31.8	32.3	54.5
Rating (r')	0	.3	0
Task (t)	.3	6.1	1.5
pr'	2.0	0	1.0
pt	14.1	34.4	16.6
tr'	.4	.2	0
ptr', e	51.5	26.7	26.4
$\widehat{\rho}^2$ (G-study)	.78	.80	.78
$\widehat{\rho}^2$ (D-study)	.68	.76	.71

Note. G-study = generalizability study; D-study = decision study.

Table 4 Scale Scores G-Study Results Including Percentage of Total Variance for Each Facet of Measurement in the Design $p \times r'$

Source	*SS*	*df*	*MS*	Variance components	%
Person (p)	2,669,751.97	1,346	1,983.47	931.86	88.6
Rating (r')	75.17	1	75.17	0	0
pr', e	16,1,174.3	1,346	119.74	119.74	11.4

Note. G-study = generalizability study.

Unexplained error and the three-way interaction between person, task, and rating (ptr', e) accounted for a relatively large percentage of total variance (51.5%) in Claim 1 performance, although ability (p) still explained a sizable percentage of the total variance (31.8%). The opposite pattern was observed in Claim 3 performance, with ability accounting for the largest percentage of total variance in performance (54.5%). Generalizability coefficients based on G-studies for each of the claim scores were reasonably high (.78 – .80), and those based on D-studies were lower (.68 – .76).

Consistency of Scale Scores

The full results of the G-study for scale scores using the $p \times r$ design are summarized in Table 4.

As seen in Table 4, in a measurement design where scale scores are portioned into variance associated with ability (p) and different sets of ratings (r'), a high percentage of the variance in scores (88.6%) is explained by ability, minimal variance is attributable to differences between scores produced by different sets of ratings (r'), and a relatively smaller percentage (11.4%) is attributable to the combination of unexplained error (e) and differences in rank ordering of test takers across the sets of ratings (pr'). The generalizability coefficient associated with the G-study design was $\widehat{\rho}^2 = .94$, and the D-study coefficient for the operational design using one set of ratings ($r' = 1$) was $\widehat{\rho}^2 = .89$.

Discussion

This study analyzed the reliability or consistency of TOEIC Speaking scores across different levels of the scoring procedure using the framework of G-theory. As expected, at the individual task level, the generalizability of scores under operational conditions varied greatly, from $\widehat{\rho}^2 = .40$ to .85. The generalizability of claim-level performances based on their constituent tasks narrowed to the range of $\widehat{\rho}^2 = .68$ (Claim 1) to $\widehat{\rho}^2 = .76$ (Claim 2) under operational conditions, coefficients that are reasonably high but do not uniformly reflect a level of score consistency that would facilitate high-stakes decisions based on performance with respect to individual claims. The generalizability of scale scores across different sets of ratings was much higher ($\widehat{\rho}^2 = .94$), and the level of consistency corresponding to operational conditions that use one set of ratings remained relatively high ($\widehat{\rho}^2 = .89$) — certainly high enough according to traditional psychometric practice to justify using these scores for high-stakes decisions. Thus this study contributes backing to several of the warrants listed in Table 1 that support the claim that TOEIC Speaking scores are consistent.

The results of the analysis of test-taker performance at the claim level provides support for the assertion that scores on different tasks within claims are internally consistent. The G-studies that examined the generalizability of claim scores found that very little of the total variance in scores could be attributed to the main effect of task controlling for rating (0.3 – 6.1%), which suggests that the overall difficulty of tasks within a claim did not vary substantially. While this study did not conduct an analysis of internal consistency in the same manner as Liao and Wei (2010), the finding that the main effect of task at the claim level does not explain a sizable proportion of total variance is evidence to support the warrant. A larger percentage of variance was explained by the interaction between ability and task ($p \times t$), which suggests that some tasks were easier or more difficult for different test takers. This could be due to differences in the nature of the tasks performed or other contextual features of tasks; regardless, the finding that task effects were comparatively larger than rating effects is consistent with prior L2 speaking research (In'nami & Koizumi, 2016).

Analyses across all three levels of the scoring procedure suggested that differences between ratings had a minimal effect on scores, which supports the claim that scores from different ratings are consistent. Most importantly, the analysis of scale scores found that minimal variance was associated with differences between scale scores for the same test taker based on sets of ratings. The generalizability coefficient associated with operational rating conditions ($\widehat{\rho}^2 = .89$) was similar in magnitude to previous research findings that used different test forms and different samples of test takers and raters to measure rater agreement using agreement rates (Liao & Wei, 2010; Qu & Ricker-Pedley, 2013) and G-theory (Liao & Wei, 2010). Although the methodological approach employed in these analyses (i.e., *rating* vs. *rater* as a facet) may lead to the underestimation of variance components associated with the rating facet, these variance component magnitudes were consistently negligible across tasks, at the claim level, and for scaled scores. Thus this series of G-studies using ratings collected under operational conditions helps strengthen the backing to support the warrant that scores from different ratings are consistent.

Thus the findings of this study provide evidence to strengthen the backing of claims about the consistency of TOEIC Speaking scores. Most crucially, the generalizability and dependability of TOEIC Speaking scale scores were found to be relatively high. While score consistency itself is not sufficient to facilitate high-quality decisions — score interpretations must be meaningful, impartial, generalizable, and relevant to those decisions (i.e., fair and valid) — this study contributes additional evidence that the psychometric basis for score interpretations is relatively strong.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford, England: Oxford University Press.

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.

Brown, A. (2012). Interlocutor and rater training. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 413 – 425). New York, NY: Routledge.

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using Edu-G*. New York, NY: Routledge.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297 – 334.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261 – 287). Mahwah, NJ: Erlbaum.

Everson, P., & Hines, S. (2010). How ETS scores the TOEIC Speaking and Writing tests responses. In D. Powers (Ed.), *TOEIC compendium* (1st., pp. 8.1 – 8.9). Princeton, NJ: Educational Testing Service.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th., pp. 65 – 110). New York, NY: American Council on Education/Praeger.

Hines, S. (2010). Evidence-centered design: The TOEIC Speaking and Writing tests. In D. Powers (Ed.), *TOEIC compendium* (1st., pp. 7.1 – 7.31). Princeton, NJ: Educational Testing Service.

In'nami, Y., & Koizumi, R. (2016). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing, 33*, 341 – 366.

Knapp, T. R., & Mueller, R. O. (2010). Reliability and validity of instruments. In G. Hancock & R. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 337 – 341). New York, NY: Routledge.

Lee, Y.-W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing, 23*, 131 – 166.

Lee, Y.-W., & Kantor, R. (2005). *Dependability of new ESL writing test scores: Evaluating prototype tasks and alternative rating schemes* (TOEFL Monograph No. MS-30). Princeton, NJ: Educational Testing Service.

Liao, C.-W., & Qu, Y. (2010). Alternate test forms test-retest reliability for the TOEIC Speaking and Writing tests. In D. Powers (Ed.), *TOEIC compendium study* (1st., pp. 11.1 – 11.40). Princeton, NJ: Educational Testing Service.

Liao, C.-W., & Wei, Y. (2010). Statistical analyses for the TOEIC Speaking and Writing pilot study. In D. Powers (Ed.), *TOEIC compendium study* (1st., pp. 9.1 – 9.25). Princeton, NJ: Educational Testing Service.

Lin, C.-K. (2013, June). *Handling sparse data in performance-based language assessments under generalizability theory framework*. Paper presented at the Language Testing Research Colloquium, Seoul, South Korea.

Qu, Y., & Ricker-Pedley, K. L. (2013). Monitoring individual rater performance for TOEIC Speaking and Writing tests. In D. Powers (Ed.), *TOEIC compendium study* (2nd., pp. 9.1 – 9.9). Princeton, NJ: Educational Testing Service.

Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. (1965). Generalizability of stratified-parallel tests. *Psychometrika, 30*, 39 – 56.

Sawaki, Y. (2017, June). *Generalizability of content analytic rating scales for assessing university-level Japanese EFL learners' summarization performance*. Paper presented at Fundamental Considerations in Language Testing: An International Conference in Honor of Lyle F. Bachman, Salt Lake City, UT.

Schmidgall, J. E. (2017). Evaluating score and decision consistency across claims in a validation argument. *Applied Measurement in Education*, *30*(4), 287 – 296. doi: 10.0180/08957347.2017.1353988

Suggested citation:

Schmidgall, J. E. (2017). *The consistency of TOEIC® Speaking scores across ratings and tasks* (Research Report No. RR-17-46). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12178

Action Editor: Donald Powers

Reviewers: Lawrence Davis and Guangming Ling