ETS TOEFL®

Monograph Series

MS - 28
March 2005

Dependability of Scores for a New ESL Speaking Test: Evaluating Prototype Tasks

Yong-Won Lee

Dependability of Scores for a New ESL Speaking Test: Evaluating Prototype Tasks

Yong-Won Lee
ETS, Princeton, NJ

RM-04-07

ETS is an Equal Opportunity/Affirmative Action Employer.

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org
Web site: www.ets.org/toefl

Foreword

The TOEFL Monograph Series features commissioned papers and reports for TOEFL 2000 and other Test of English as a Foreign Language™ (TOEFL®) test development efforts. As part of the foundation for the development of the next generation TOEFL test, papers and research reports were commissioned from experts within the fields of measurement, language teaching, and testing through the TOEFL 2000 project. The resulting critical reviews, expert opinions, and research results have helped to inform TOEFL program development efforts with respect to test construct, test user needs, and test delivery. Opinions expressed in these papers are those of the authors and do not necessarily reflect the views or intentions of the TOEFL program.

These monographs are also of general scholarly interest, and the TOEFL program is pleased to make them available to colleagues in the fields of language teaching and testing and international student admissions in higher education.

The TOEFL 2000 project was a broad effort under which language testing at Educational Testing Service® (ETS®) would evolve into the 21st century. As a first step, the TOEFL program revised the Test of Spoken English™ (TSE®) and introduced a computer-based version of the TOEFL test. The revised TSE test, introduced in July 1995, is based on an underlying construct of communicative language ability and represents a process approach to test validation. The computer-based TOEFL test, introduced in 1998, took advantage of new forms of assessment and improved services made possible by computer-based testing, while also moving the program toward its longer-range goals, which included:

- the development of a conceptual framework that takes into account models of communicative competence
- a research program that informs and supports this emerging framework
- a better understanding of the kinds of information test users need and want from the TOEFL test
- a better understanding of the technological capabilities for delivery of TOEFL tests into the next century

Monographs 16 through 20 were the working papers that laid out the TOEFL 2000 conceptual frameworks with their accompanying research agendas. The initial framework document, Monograph 16, described the process by which the project was to move from identifying the test domain to building an empirically based interpretation of test scores. The subsequent framework documents, Monographs 17-20, extended the conceptual frameworks to the domains of reading, writing, listening, and speaking (both as independent and interdependent domains). These conceptual frameworks guided the research and prototyping studies described in subsequent monographs that resulted in the final test model. The culmination of the TOEFL 2000 project is the next generation TOEFL test that will be released in September 2005.

As TOEFL 2000 projects are completed, monographs and research reports will continue to be released and public review of project work invited.

TOEFL Program
Educational Testing Service

Abstract

A new multitask speaking measure is expected to be an important component of a new version of the Test of English as a Foreign Language™ (TOEFL®). This study considered two critical issues concerning score dependability of the new speaking measure: How much would the score dependability be impacted by (a) combining scores on different task types into a composite score and (b) rating each task only once? To answer these questions, the study used generalizability theory (G-theory) procedures to examine (a) the relative effects of tasks and raters on examinees' speaking scores and (b) the impact of the numbers of tasks and raters per speech sample and of subsection lengths on the dependability of speaking section scores. Univariate and multivariate G-theory analyses were conducted on rating data collected for 261 examinees for the study. The finding in the univariate analyses was that it would be more efficient to increase the number of tasks rather than the number of ratings per speech sample in maximizing the score dependability. The multivariate G-theory analyses also revealed that (a) the universe scores among the task-type subsections were very highly correlated and that (b) slightly larger gains in composite score reliability would result from increasing the number of listening-speaking tasks for the fixed section lengths.

Key words: Dependability index, EFL/ESL (English as a foreign/second language), generalizability coefficients, generalizability theory, independent tasks, integrated tasks, rating design, score dependability, speaking assessment, task generalizability, variance components

Acknowledgments

This research project was funded by the TOEFL (Test of English as a Foreign Language) Program at ETS. Several members of the ETS staff and external reviewers, in addition to the author, contributed to this project. Fred Cline prepared data sets for this analysis, and Youn-Hee Lim assisted in creating tables and graphs. Brent Bridgeman, Craig Deville, Antony Kunnan, Phil Everson, and Don Powers reviewed preliminary manuscripts and provided helpful comments for preparing the draft report. I also would like to thank Dan Eignor, Mary Enright, Tim McNamara, Pamela Mollaun, Deanna Morgan, and Hariharan Swaminathan for their review and constructive feedback about earlier versions of this report. Needless to say, the responsibility for any errors that remain are solely the author's, and the ideas and opinions expressed in this report are those of the author, not necessarily of ETS or the TOEFL Program.

Table of Contents

List of Tables

Page

List of Figures

Page

Introduction

A new multitask speaking measure is expected to be an important component of a new version of the Test of English as a Foreign Language™ (TOEFL®), as first envisioned in the *TOEFL® 2000 Speaking Framework* (Butler, Eignor, Jones, McNamara, & Suomi, 2000). Three major types of speaking tasks have been considered for the speaking section of this new test: *independent* speaking tasks (e.g., tasks based on a stand-alone statement or visual prompt) and two types of *integrated* tasks (listening-speaking and reading-speaking). Independent tasks require the examinees to use their personal experiences or general knowledge to respond to a speaking task, whereas integrated tasks require examinees first to understand academic lectures or texts and then to prepare spoken responses that demonstrate understanding of such stimulus material. However, assessments that require test-takers to provide extended constructed responses often lack score generalizability across tasks or task types (Breland, Bridgeman, & Fowles, 1999; Cumming, Kantor, Powers, Santos, & Taylor, 2000; Miller & Linn, 2000) and depend on subjective rater (or judge) judgment for scoring the examinee responses.

For a similar reason, task and rater variabilities have been researched as two major sources of measurement error in the context of performance-based language assessment (Bachman, Lynch, & Mason, 1995; Breland et al., 1999; Brennan, Gao, & Colton, 1995; Fulcher, 2003; Henning, 1996; Lynch & McNamara, 1998; van Weeren & Theunissen, 1987). In terms of tasks, different types of tasks are associated with different types of input stimuli (e.g., a lecture, a reading passage, a stand-alone written or visual prompt) in the new speaking assessment. Thus, one intriguing research question is whether examinees' performance on one task would be very similar to their performance on other tasks designed to measure a common construct of interest (i.e., speaking proficiency). Potentially, each of these task types might be tapping a somewhat distinct aspect of speaking and, if the speaking scores were based on a set of these heterogeneous task types, the reliability of the composite scores would be negatively impacted (or the impact of other skills may be confounded with the speaking scores). In that respect, it is very important to examine the generalizability of speaking scores across tasks and task types in evaluating and validating a new speaking measure.

Another major factor that might affect score reliability is raters (or ratings). These tasks are scored by raters who do not always agree. Therefore, score variability attributable to ratings needs to be closely examined. Given that a speaking assessment consisting of multiple tasks is

expected to be an important component of the new TOEFL examination, the number of performance samples to be rated would likely be drastically larger than for the other tests in the TOEFL family of performance-based assessments, which are taken by a much smaller number of examinees than TOEFL. Moreover, if each speech sample in the new speaking section were to be rated twice, as is done for other TOEFL tests, the task of processing and rating speech samples would be even more challenging for the TOEFL Program. An alternative rating design would be to use a single rating per speech sample, preferably with a detection mechanism to flag an unusual rating (that is extremely higher or lower than expected) for adjudication. This option is attractive for the TOEFL Program because it would be more economical.

Two critical questions then are how much the score dependability would be affected by (a) combining scores on different task types into a composite score and (b) rating each task only once. One purpose of this study was, therefore, to examine the relative effects of tasks and raters on examinees' total speaking scores based on integrated and independent tasks, and the impact of the number of tasks and raters on the score dependability in the generalizability theory (G-theory) framework. A second purpose was to determine, through a multivariate G-theory analysis, the optimal configuration of task types and number of speaking tasks to maximize the reliability of the composite score for fixed section lengths.

Integrated and Independent Tasks in Speaking Assessment

As previously mentioned, both integrated and independent speaking tasks have been considered as possible candidates for assessing speaking in the new TOEFL assessment (Butler et al., 2000). These task types are intended to elicit responses that reflect spoken skills needed in an academic environment. Integrated tasks require examinees to integrate multiple language skills in a substantial way to complete a speaking task at hand (e.g., to understand academic texts and lectures and then create spoken responses that demonstrate understanding of the texts and lectures). While the integrated tasks provide the information about which examinees will speak, the independent tasks usually require examinees to rely on their personal experiences or general knowledge to complete the task. Types of integrated tasks to be included in the new speaking assessment would likely require integration of two language skills, as in listening-speaking and reading-speaking tasks. In contrast, independent speaking tasks may be similar to tape-mediated TSE® (The Test of Spoken English™) tasks, which are based on stand-alone visuals or

statements. Integrated tasks have been advocated for two main reasons (Lewkowicz, 1997): (a) Test takers are less likely to be disadvantaged by insufficient information upon which to base their argument (Read, 1990; Weir, 1993) and (b) validity would be enhanced by simulating real-life communication tasks in academic contexts (Wesche, 1987).

Some concerns can be raised, however, about speaking assessments composed of integrated tasks (e.g., task generalizability, dependency across test sections). These concerns may be also related to the role of input stimuli in eliciting examinees' spoken responses. A claim might be also made that each of these different speaking task types measure a somewhat distinct construct and, therefore, separate scores should be reported for each of these distinct constructs. Because the three task types being considered for the new TOEFL speaking section (independent tasks and the two types of integrated tasks) are dissimilar in input stimuli characteristics (i.e., a stand-alone written statement or visual prompt, a reading passage, an auditory text), test takers might use different cognitive skills and processes in responding to them. A similar argument could be made about the rating process for examinees' speech samples. Raters are expected to apply somewhat different scoring criteria for different task types. When rating examinee responses for independent tasks, for example, raters can mostly focus on language. When rating examinee responses from integrated tasks, however, raters also have to attend to content accuracy to make sure that the examinees have adequately understood the information that has been presented in the text or lecture.

However, if the seemingly distinct constructs associated with these three task types can be shown to be highly correlated from a psychometric viewpoint, reporting a composite score for these task types would be justifiable. In that sense, it remains to be seen whether the different types of tasks can be shown to be truly additive in terms of the speaking construct they are intended to measure as a whole. Whether the three task scores can be aggregated to form a single, reliable speaking score (or a single composite) can be viewed as an empirical question.

Investigation of Score Dependability: Generalizability Theory

When only a single measurement facet is involved in the assessment system, classical test theory (CTT) is sufficient for examining the generalizability of test scores from a norm-referenced testing perspective, as exemplified by internal consistency reliabilities. Speaking, however, involves more than one major random facet. These facets include, at least, tasks and raters as major sources of score variability. Such a context clearly requires employing a

multifaceted analysis—generalizability theory (G-theory; Cronbach, Gleser, Nanda, & Rajaratnam, 1972)—that can analyze more than one measurement facet simultaneously, in addition to the object of measurement (i.e., examinees).[1]

Univariate and Multivariate Generalizability Theory

G-theory provides a comprehensive conceptual framework and methodology for analyzing more than one measurement facet in investigations of assessment error and score dependability (Brennan, 1992, 2000, 2001; Cronbach et al., 1972; Shavelson & Webb, 1991; Suen, 1990). Through a two-staged investigation that includes a generalizability study (G-study) and a decision study (D-study), G-theory enables researchers to disentangle multiple sources of measurement error and investigate the impact of various changes in the measurement design on score reliabilities. In the G-study, the variances associated with different facets of measurement—including the object of measurement (usually examinees)—are estimated and evaluated in terms of their relative importance in contributing to the total score variance, given a universe of admissible observations (Brennan, 2001). In the D-study, the impact of various changes in the measurement design (e.g., different numbers of tasks or raters, standardization of rating procedures) on score reliability is investigated for the universe of generalization of interest (Brennan, 2001).

In the D-study, two different types of reliability coefficients can be computed, one for norm-referenced and the other for criterion-referenced score interpretations, respectively: (a) a generalizability coefficient ($E\rho^2$ or G) and (b) a dependability index (Φ or Phi). A generalizability coefficient that uses relative error variance [$\sigma^2(\delta)$] as error variance can be conceptualized as the ratio of universe (true) score variance to expected observed score variance (Brennan, 2001; Cronbach et al., 1972). It is also analogous to a reliability coefficient (i.e., coefficient alpha) in classical test theory, but a classical reliability coefficient usually implies a single undifferentiated source of measurement error. To emphasize the multifaceted nature of measurement error, the term *generalizability coefficient* is used to describe the reliability coefficient obtained in the D-study for norm-referenced score interpretation (Suen, 1990). In contrast, a dependability index uses absolute error variance [$\sigma^2(\Delta)$] as error variance and is more appropriate for domain-referenced or criterion-referenced situations (Brennan, 2001). The generalizability coefficients are useful in testing situations where the purpose of measurement is

to make relative decisions about examinees (e.g., selection of individuals for a particular program) based on the relative standing (or rank ordering) of examinees compared to others in the same group or to a group average in test scores (Bachman, 1990; Bolus, Hinofotis, & Bailey, 1982; Shavelson & Webb, 1991). However, when the measurement objective is to make absolute decisions about whether examinees have attained a prespecified criterion level of performance, it is more appropriate to use the reliability coefficient (e.g., Φ) that takes into account such systematic differences related to test forms, tasks, and raters.[2]

Univariate G-theory can be also extended for the multivariate situations where a test is made up of multiple subsections or subtests, and there is a need to investigate the reliability of the composite of subsection scores as well as each subsection score (Brennan, 2001). In the multivariate G-theory design, a set of subsections or content categories in the test is considered a fixed facet, and the number of elements in each fixed content category can be either the same (balanced) or different (unbalanced) across the categories. Both univariate and multivariate G-theories have been developed within the same G-theory framework, but differ in that the former is suited to analyzing scores from one test at a time, whereas the latter can also handle scores from multiple subsections simultaneously (see Brennan, 2001). Another quite often cited use of multivariate G-theory analyses is to analyze a test for which multiple test forms are developed according to the same table of specifications. In this type of test, the same specification, such as structure of content categories for items, is applied across different forms of the test, and thus content categories can be safely regarded as a fixed facet.

In the context of the new TOEFL assessment, task types in the speaking section (i.e., listening-speaking, reading-speaking, independent speaking) can be viewed as a fixed-content facet for multivariate G-theory analyses because all three task types are likely to appear repeatedly in each form of the speaking assessment according to the test specifications. Thus, these three task types can be treated as a fixed-content category facet in the multivariate G-theory framework so that both the univariate and multivariate analyses can be conducted on the same data. If we are simply interested in examining the impact on the score reliability of different numbers of tasks and raters in the whole section, ignoring the task-type subsection boundary, we can just use univariate G-theory to estimate variance components and score reliability coefficients for the total section. However, if we are interested in what combination of subsection lengths for a fixed total section length can maximize the composite score reliability of

the section, the multivariate G-theory analysis can provide answers to such questions. More importantly, the universe score correlations among the subsections estimated in multivariate G-theory analyses can provide a basis for determining whether these subsection scores can sensibly be combined into a single composite score for the whole section.

G-theory Application in Large-Scale Assessments

When multiple raters (r) and tasks (t) are involved in the assessment of proficiency of examinees or persons (p), the most powerful G-study design from the researcher's perspective is a fully crossed, two-facet design with tasks and raters as random facets ($p \times t \times r$). This requires that all tasks be rated by all raters for all examinees in the data-collection stage for G-theory analyses. Once data are collected according to such a fully crossed design, researchers can investigate the score reliability for various nested as well as crossed assessment scenarios in the D-study (Brennan, 2001). When such a data-collection scheme is employed in large-scale, performance-based assessments, however, each rater is required to rate an unrealistically large number of performance samples on multiple tasks for examinees in a single rating session. Thus, it would not be feasible to collect rating data using this design in an operational test setting unless, for research purposes, a smaller subsample of examinees is rated or rerated according to this ideal design in a special rating session.

For this reason, some variations of a partially nested rating design are generally used for scoring performance samples in many large-scale, performance-based assessments. Often two raters are assigned to rate each examinee on a single task (or multiple tasks), and their ratings are averaged to compute the reported score. In this scenario, different pairs of trained raters are chosen from a pool to rate different examinees within each task. The raters are usually allowed to rate examinee speech samples across tasks. This scenario was used in the rating design for prototyping the new speaking tasks used in this study. The closest G-study design representing such a testing situation is a partially nested G-study design, with tasks (t) and raters (r) as two random facets $[(r:p) \times t]$.[3] This design may be used with caution to investigate the joint impact of the number of tasks and raters on score dependability in such a context. It should be pointed out, however, that the G-theory analyses based on the nested design may not be applicable for such contexts in the strictest sense, because some degree of rater overlap is usually allowed across

examinees or blocks of examinees in operational testing situations (including the rating design used in this study).

An alternative G-study design for such a situation would be to treat ratings (i.e., the first and second rating: r′), instead of raters (r), as a random facet. Since all examinees' final scores are based on the two ratings, it would be possible to use a fully crossed, two-facet design with two random facets of tasks and ratings ($p \times t \times r'$), if we use the ratings rather than raters as a measurement facet. This alternative strategy is also consistent with the inter-rater reliability computation procedure used for large-scale, performance-type language assessments, such as the Test of Spoken English (TSE), the Test of Written English™ (TWE®), and the computer-based TOEFL (TOEFL-CBT) essay test, where the inter-rater reliability is computed by adjusting the correlation between the first and second ratings for the number of ratings per performance sample. In other words, the ratings (not raters) are used as a unit of analyses to obtain the inter-rater reliability estimate. The same approach has been used by researchers in language assessment as an alternative to, or together with, a partially nested design (Bachman et al., 1995; Lee, Golub-Smith, Payton, & Carey, 2001, Lee & Kantor, in press). Recently, Lee and Kantor (in press) found in their G-theory analyses of original ($p \times t \times r'$) and rerated ($p \times t \times r$) writing data that both the $p \times t \times r'$ and $p \times t \times r$ designs resulted in very similar estimates of the overall proportion of rater-related error variances and the score reliabilities for single- and double-rating situations for different test lengths. The results of the $p \times t \times r'$ design can also be compared to those of the $(r:p) \times t$ design, based on the same data for reasonableness.

One advantage of the $p \times t \times r'$ design in the context of the current study is that its multivariate counterpart ($p\bullet \times t^{o} \times r'\bullet$) is feasible in the currently available computer program for multivariate G-theory analysis, mGENOVA (Brennan, 1999), whereas the multivariate counterpart of the partially nested design is not.[4] Given the technical capacity of the estimation software, the only designs of interest that are amenable for such analyses in the current study are the multivariate counterparts of the two crossed designs of $p \times t \times r'$ and $p \times t$.

Research Questions

A special focus was placed on investigating the impact of the number of tasks (varied from 1 to 12) and of the number of ratings (1 or 2). Both univariate and multivariate analyses were designed, with the following four questions in mind:

1. What would be the impact of increasing the number of tasks from 1 to 12?
2. What would be the impact of increasing the number of ratings per essay from 1 to 2?
3. Are the universe (or true) score correlations among the three speaking subsections high enough to justify combining them into a single composite score?
4. What combinations of task-type subsection lengths for fixed total lengths (e.g., 5 tasks) would maximize the composite score reliability for speaking?

Method

Participants

Participants were English as a second language (ESL) students recruited from three domestic and five international (Australia, Canada, Hong Kong, Mexico, and Taiwan) testing sites. Most of the participants from English-speaking countries (e.g., domestic sites, Australia, and Canada) were ESL students enrolled in either regular academic degree programs at either graduate or undergraduate levels or intensive English language training programs at colleges/universities in each country. Participants completed a battery of English assessments containing a prototype version of speaking task types in the autumn of 2000 (Enright et al., 2003). Ratable speaking responses on some or all of the tasks were available for 478 examinees. However, only the data from the 261 participants who had scorable responses for all of the 11 tasks used in this study were analyzed. Of these 261 participants, 141 were male, 113 were female, and 7 were of unidentified gender. At the time of testing, the average age of the examinees was approximately 24 years old. Their paper-based TOEFL Institutional Testing Program (ITP) scores ranged from 337 to 673, with a mean of 577 and a standard deviation of 57. The participants were from 31 diverse native language backgrounds, with the four largest native language groups being Spanish (31%), Chinese (29%), Korean (8%), and Thai (5%).

Instrument

A total of 13 speaking tasks for the three task types were prepared and administered originally as part of a prototyping study (Enright et al., 2003). These speaking tasks included 6 listening-speaking (LS), 2 reading-speaking (RS), and 5 independent speaking (IS) tasks (see Appendix C for examples of each task type). Two LS tasks that were based on academic

conversations were excluded from the analysis, however, because these tasks did not provide enough substance for spoken responses. For this reason, only 11 tasks (4 LS, 2 RS, 5 IS) taken by the 261 examinees were analyzed in the study.

Rating Procedures

Two distinct scoring rubrics were used for the integrated and independent tasks. Each examinee response was double-rated on a scale of 1 to 5 (see Appendix D). Different pairs of independent raters was selected from a pool of 22 raters and assigned to each speech sample. Raters had a chance to rate all the tasks in the test (i.e., rater overlap was allowed across tasks), but raters were nested within examinees (with some degree of rater overlap across examinees). To minimize the potential halo effect, they were also asked to rate all the speech samples for a specific task for all examinees before moving on to the next task within a particular task type. Once all the tasks within the particular task type were rated, raters moved to the next task type and the process was repeated.

Data Analysis

The computer program GENOVA (Crick & Brennan, 1983) was used to estimate the variance components and the score reliability coefficients (e.g., $E\rho^2$, Φ) in the univariate analysis. The computer program mGENOVA (Brennan, 1999) was used for the multivariate analyses to estimate the variance and covariance components and the reliability coefficients for the subsections and composite scores.

For the univariate analysis, three G-study designs were used for the speaking data to estimate the variance components in the G-study: (a) a two-facet, partially nested design [(r:p) × t] with tasks (t) and raters (r) as random facets; (b) a two-facet crossed design (p × t × r′) with tasks (t) and ratings (r′) as random facets; and (c) a single-facet crossed design (p × t) with tasks (t) as random facets that used averaged ratings over two raters as the unit of analysis. The first two were the two main comparison G-study designs used in this study to investigate the relative effects of tasks and raters together (see also the previous section, "Investigation of Score Dependability: Generalizability Theory," for rationales for these two designs). However, the third design (p × t) was used to estimate internal consistency reliability coefficients (α_T) for different section lengths when the averaged ratings over two raters were used as units of analysis; thus, possible scores were 1.0, 1.5, …, 4.5, 5.0. In a single-facet design, a Cronbach

alpha (α) is numerically equivalent to a generalizability ($E\rho^2$) coefficient (Brennan, 1992, 2001; Suen, 1990). D-studies were designed for the same universe of generalizations as the universe of admissible observations used in the G-study for all three designs [i.e., (R:p) × T, p × T × R′, p × T]. For the first two designs, multiple D-studies were carried out by varying both the number of tasks from 1 to 12 and the number of ratings from 1 to 2. For the third design (p × t), only the number of tasks could be manipulated because the averaged ratings over two raters were used as units of analysis to estimate the variance components. In addition, standard errors of measurement (SEMs) were computed based on the relative and absolute error variances [$\sigma^2(\delta)$, $\sigma^2(\Delta)$] for various testing scenarios, and confidence intervals based on the absolute SEM [$\sigma(\Delta)$] for a universe score of 3 for different section lengths were constructed for both the single- and double-rating situations for comparison.

For the multivariate analysis, both the two-facet and single-facet crossed designs were used to estimate the variance and covariance components in the G-study for the task type subsections (i.e., LS, RS, and IS) of the speaking section: (a) a two-facet crossed design with tasks and ratings as random facets ($p\bullet \times t^o \times r'\bullet$); (b) a single-facet crossed design with tasks as a random facet ($p\bullet \times t^o$). In the first design ($p\bullet \times t^o \times r'\bullet$), it was assumed that the persons (p) and ratings (r′) were crossed with the LS, RS, and IS subsections (v), but tasks (t) were nested within each subsection (v). In the second design ($p\bullet \times t^o$), however, the examinees were crossed with the subsection, but the tasks were nested within the subsections. In this design, the averaged ratings over two raters were also used as units of analysis in each subsection, as in its univariate counterpart (p × t). Multiple D-studies were carried out by varying the number of tasks in two major ways: (a) by increasing the number of tasks in each subsection simultaneously and (b) by manipulating the number of tasks in each subsection for the several fixed total section lengths of interest.

Of particular interest were comparisons of composite score reliabilities for different combinations of subsection lengths for the fixed total section lengths of 3, 4, and 5 tasks. When the total section length was 3 tasks, the only possible scenario for representing all three subsections was to take 1 task for each of the LS, RS, and IS subsections (1-1-1). For the test length of 4 tasks, there were three possible scenarios (1-1-2, 1-2-1, and 2-1-1). When the total test length was 5 tasks, there were six possible scenarios (1-1-3, 1-2-2, 1-3-1, 2-2-1, 3-1-1, and

2-1-2). For comparison purposes, two additional combinations for the total test length of 6 that were of interest to the test development team were included in the D-study, along with the two longer test scenarios of 11 and 12 tasks (4-2-5, 4-3-5).

Results

Similar results were obtained about the impact of the numbers of tasks and raters on the score reliability in both univariate and multivariate analyses. In the univariate analyses, both the $p \times t \times r'$ and $(r:p) \times t$ designs yielded almost identical results (e.g., relative proportions of various variance components and score reliabilities for different numbers of tasks and raters). Multivariate analyses also provided useful information about the relationships between the task-type subsections. More detailed descriptions of the results of univariate and multivariate analyses are presented in the next section.

Univariate Analysis [(r:p) × t, p × t × r', p × t]

Estimated variance components. Table 1 displays the estimated G-study variance components, standard errors of estimated variances (S.E.), and percentage of each variance component contributing to the total score variance in each of the $(r:p) \times t$ and $p \times t \times r'$ designs.

A total of five variance components were estimated for the $(r:p) \times t$ design in the G-study, which included the variance components associated with the examinee [$\sigma^2(p)$], task [$\sigma^2(t)$], examinee-by-task interaction [$\sigma^2(pt)$], rater (nested within examinees) [$\sigma^2(r:p)$], and task-by-rater (nested within examinee) interaction plus undifferentiated error [$\sigma^2(tr:p, \text{undifferentiated})$] effects (see Table 1). Among them, the largest variance component was that associated with examinees [$\sigma^2(p)$], which explained about 51.3% of the total variance estimated for a single observation in the G-study. The second largest variance component was the one associated with the interaction between tasks and raters nested within examinees plus undifferentiated error [$\sigma^2(tr:p, \text{undifferentiated})$], which accounted for about 27.9% of the total variance in the G-study. The third largest variance component was that for the examinee-by-task interaction effect [$\sigma^2(pt)$], which accounted for about 17.1% of the total variance. This indicates that a significant portion of examinees were not rank-ordered consistently across different tasks. Nevertheless, the variance components associated with the main effects of tasks [$\sigma^2(t)$] and raters (nested within examinees) [$\sigma^2(r:p)$] were very small, explaining about 1.8% of the total variance each. The

small task variance indicates that tasks used in this study were not much different in difficulty overall, whereas the small raters (nested within examinees) effect suggests that the confounded effects of rater severity differences and rater inconsistency across examinees were very small in this study.

Table 1

Estimated Variance Components for G- and D-studies in the Univariate Analyses for the New TOEFL Speaking Section, Based on the Prototyping Study Data ($n_p = 261$, $n_t = 11$, $n_{r:p} = 2$; $n_p = 261$, $n_t = 11$, $n_{r'} = 2$)

Effects	G-study [(r:p) × t]			Effects	G-study (p × t × r′)		
	Single observation				Single observation		
	Variance	S.E.	Percent		Variance	S.E.	Percent
Examinee (p)	0.669	0.063	51.3	Examinee (p)	0.669	0.063	51.3
Task (t)	0.024	0.010	1.8	Task (t)	0.022	0.010	1.7
Rater nested within examinees (r:p)	0.023	0.005	1.8	Rating (r′)	0.000	0.000	0.0
Examinee-by-task (pt)	0.223	0.012	17.1	Examinee-by-task (pt)	0.225	0.012	17.3
Task-by-rater-nested within examinees (tr:p, undifferentiated)	0.364	0.010	27.9	Examinee-by-rating (pr′)	0.024	0.005	1.8
				Task-by-rating (tr′)	0.003	0.002	0.3
				Examinee-by-task-by-rating (ptr′, undifferentiated)	0.360	0.010	27.6
Total	1.302		100.0	Total	1.302		100.0

Table 1 also shows the seven variance components estimated for the p × t × r′ design in the G-study, which included the variance components associated with the examinee [σ^2(p)], task [σ^2(t)], rating [σ^2(r′)], examinee-by-task interaction [σ^2(pt)], examinee-by-rating interaction [σ^2(pr′)], task-by-rating interaction [σ^2(tr′)], and examinee-by-task-by-rating interaction plus undifferentiated error [σ^2(ptr′, undifferentiated)] effects. Of the seven G-study variance components, the largest variance component was again the σ^2(p) component, as in the (r:p) × t design, which explained about 51.3 % of the total variance in the G-study. The second largest variance component was that for the examinee-by-task-by-rating interaction plus undifferentiated error [σ^2(ptr′, undifferentiated)], which accounted for about 27.6% of the total variance in the G-study. The third largest interaction component was that for the examinee-by-task interaction variance (i.e., explaining about 17.3% of the total variance), suggesting that a significant portion of examinees were not rank-ordered consistently across tasks. In contrast, the relative effect of the examinee-by-rating interaction variance [σ^2(pr′)] component was very small (i.e., explaining only 1.8% of the total variance), indicating that the rank-ordering of examinees was relatively consistent across the first and second ratings. The variance component for the main effect for tasks [σ^2(t)] explained only 1.7% of the total variance, which means that tasks used in this study varied only slightly in difficulty. However, both the variances for the main effect for ratings [σ^2(r′)] and the task-by-rating interaction effect [σ^2(tr′)] were nearly zero.

Interestingly enough, the sizes of the examinee variance from the (r:p) × t and p × t × r′ designs were identical when they were rounded off to the third decimal point (as shown in Table 1). The *task* and *examinee-by-task interaction* variance values from these two designs were also very similar. It should be noted that the *examinee* variance becomes a *universe (true) score* variance later in the D-study, while the remaining variance components (other than the examinee variance) are used to define *relative* and *absolute error* variances. Even though they were not shown in Table 1, the rounded values of these relative and absolute error variances for comparable D-study assessment scenarios were also identical in the (R:*p*) × T and *p* × T × R′ designs.

Estimated reliability coefficients, SEM, and confidence intervals. Table 2 and Figure 1 show the reliability coefficients ($E\rho^2$, Φ, α_T) estimated in the D-study for the *p* × T × R′ and *p* × T designs. Table 3 and Figure 2 display the standard errors of measurement (SEMs) estimated in the D-study for these two designs. Since the estimates of the reliability coefficients

from the (R:p) × T and p × T × R′ designs were numerically identical when rounded off to a second decimal point, only those from the latter design are reported here. In addition, two highlighted assessment designs for comparison were scenarios of 5 speaking tasks rated once and 12 tasks rated twice.[5] The estimated reliability coefficients and SEM are discussed, with a focus on these two scenarios. Figure 3 shows graphically the confidence intervals for the universe score of 3 based on the absolute SEM [$\sigma(\Delta)$] estimated for the p × T × R′ design for single- and double-rating situations to illustrate the impact of single rating on score dependability on the score metric.

Table 2

Estimated Reliability Coefficients for the New TOEFL Speaking Section, Based on the Prototyping Study Data

No. of tasks	p × T × R′				p × T
	One rating per speech sample		Two ratings per speech sample		Averaged ratings
	$E\rho^2$	Φ	$E\rho^2$	Φ	α_T (or $E\rho^2$)
1	0.52	0.51	0.62	0.60	0.63
2	0.68	0.67	0.76	0.75	0.77
3	0.75	0.75	0.82	0.81	0.83
4	0.80	0.79	0.86	0.85	0.87
5	0.83	0.82	0.88	0.88	0.89
6	0.85	0.84	0.89	0.89	0.91
7	0.86	0.86	0.91	0.90	0.92
8	0.87	0.87	0.91	0.91	0.93
9	0.88	0.88	0.92	0.92	0.94
10	0.89	0.89	0.93	0.92	0.94
11	0.90	0.90	0.93	0.93	0.95
12	0.90	0.90	0.94	0.93	0.95

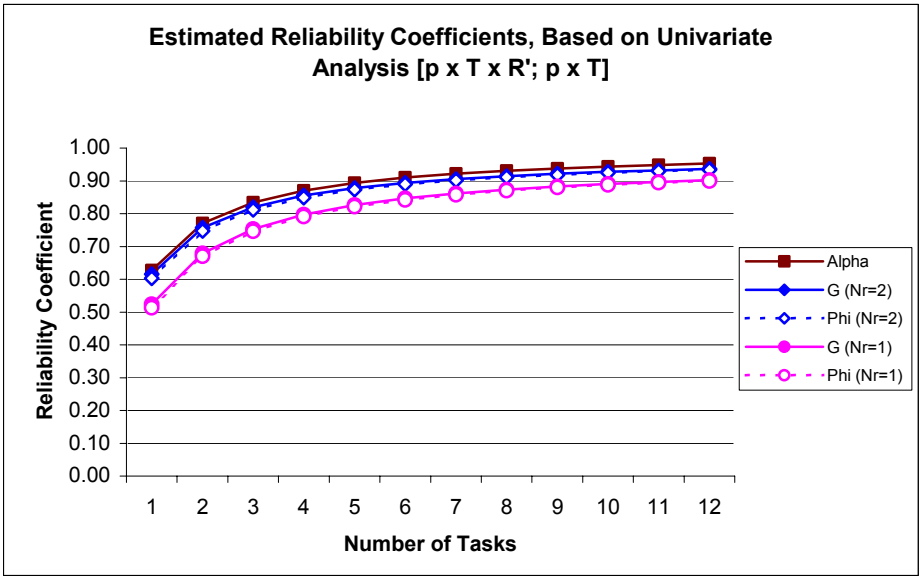


***Figure 1*. Reliability coefficients for one and two ratings per speech sample scenarios for different section lengths.**

As shown in Table 2 and Figure 1, the impact of increasing the number of tasks on the score reliability was relatively large, but the relative impact of the number of ratings per speech sample on the score reliability was very small. When the number of tasks was increased from 1 to 5 in the single-rating scenario, there was a drastic 0.31 increase (from 0.51 to 0.82) in the dependability index (Φ). An additional 0.08 increase (to 0.90) in the Φ-index occurred when the number of tasks was further increased from 5 to 12. In contrast, there was only a 0.06 increase (from 0.82 to 0.88) in the Φ-index for the 5-speaking-tasks test, when the number of ratings was increased from 1 to 2. Increases in the Φ-index value (due to the adoption of a double- rather than a single-rating scheme) ranged from 0.03 to 0.09 for various section lengths of 1 to 12 tasks. As the section length became longer, the increase in reliability due to the double-rating scheme became smaller. A similar pattern of impact was observed for the generalizability coefficient ($E\rho^2$), with the $E\rho^2$ coefficients being higher than the Φ-indices for all the assessment scenarios, as expected. By definition, these $E\rho^2$ coefficients should be at least as large as the Φ-indices (see Appendix A for more details).

Even when comparisons were made between the single- and double-rating testing scenarios with the total number of ratings per examinee in the test being held constant, increasing the number of tasks also turned out to have a larger impact on score reliability than did increasing the number of ratings per speech sample. For the total ratings of 2 per examinee in the test, for instance, equivalent single- and double-rating scenarios for comparison in the $p \times \mathrm{T} \times \mathrm{R}'$ design were: (a) 2-tasks-single-rated and (b) 1-task-double-rated scenarios. As shown in Table 3, the first scenario produced a higher Φ-index (0.67) than the second one (0.60). The same trend was observed for the total ratings of 4 (4-tasks-single-rated vs. 2-tasks-double-rated), 6 (6-tasks-single-rated vs. 3-tasks-double-rated), 8 (8-tasks-single-rated vs. 4-tasks-double-rated), 10 (10-tasks-single-rated vs. 5-tasks-double-rated), and 12 (12-tasks-single-rated vs. 6-tasks-double-rated). A similar pattern was observed for the $E\rho^2$ coefficients.

Table 3

Estimated Standard Error of Measurement for the New TOEFL Speaking Section, Based on the Prototyping Study Data

No. of tasks	$p \times \mathrm{T} \times \mathrm{R}'$				$p \times \mathrm{T}$
	One rating per speech sample		Two ratings per speech sample		Averaged ratings
	$\sigma(\delta)$	$\sigma(\Delta)$	$\sigma(\delta)$	$\sigma(\Delta)$	$\sigma(E)$
1	0.78	0.80	0.65	0.66	0.64
2	0.56	0.57	0.46	0.48	0.45
3	0.47	0.48	0.38	0.39	0.37
4	0.41	0.42	0.34	0.34	0.32
5	0.38	0.38	0.30	0.31	0.28
6	0.35	0.35	0.28	0.29	0.26
7	0.33	0.33	0.26	0.27	0.24
8	0.31	0.31	0.25	0.26	0.22
9	0.30	0.30	0.24	0.24	0.21
10	0.29	0.29	0.23	0.23	0.20
11	0.28	0.28	0.22	0.23	0.19
12	0.27	0.27	0.21	0.22	0.18

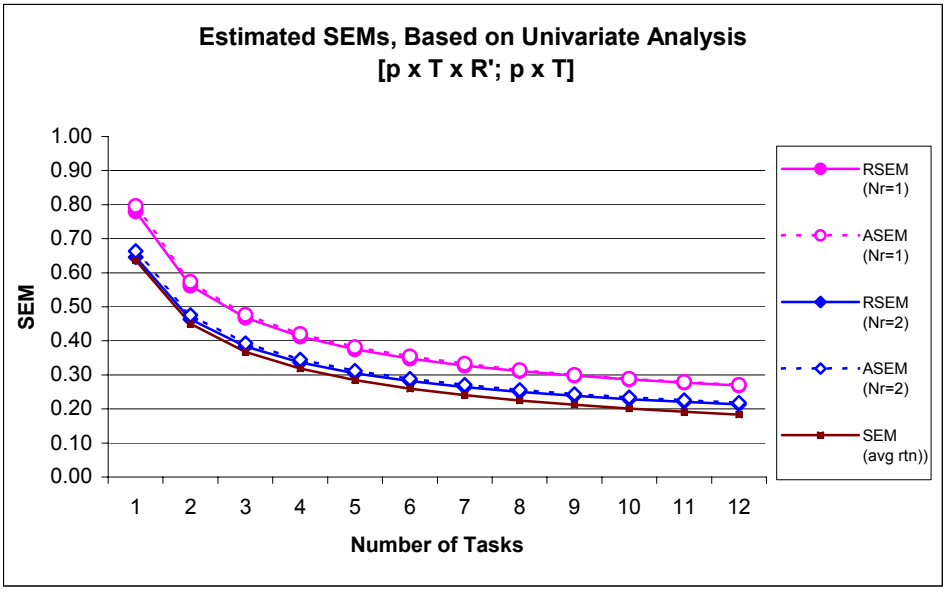


Figure 2. **Estimated standard errors of measurement (SEM) for one and two ratings per speech sample scenarios for different section lengths.**

The internal consistency alpha (α_T) coefficient also increased drastically from 0.63 to 0.89 (a 0.26 increase) when the number of tasks increased from 1 to 5, but the increase tended to taper off (only a 0.06 increase) when the number of tasks was further increased from 5 to 12 (see Table 2). The coefficient α_T from the $p \times T$ design based on the averaged ratings over two raters is usually expected to be higher than the $E\rho^2$ coefficient estimated from the $p \times T \times R'$ design for the double-rating situation, because ratings (r′) are treated as a hidden fixed facet in the $p \times T$ design and because the variances associated with ratings become part of the universe score variance in the $p \times T$ design.[6] The internal consistency alpha based on averaged ratings was slightly higher than the $E\rho^2$ coefficient for the scenario of two ratings per speech sample, as expected, but it was very close to the $E\rho^2$ coefficient.

Table 3 shows the standard errors of measurement (SEMs) estimated in the D-study for the $p \times T \times R'$ and $p \times T$ designs. The estimated relative [$\sigma(\delta)$, $\sigma(E)$] and absolute [$\sigma(\Delta)$] SEMs for various section lengths are also visually displayed in Figure 2. As shown in Table 3 and Figure 2, the decrease in SEM was much larger when the number of tasks was increased from 1 to 5 than when the number of tasks was increased from 5 to 12. When the number of tasks was increased from 1 to 5, there was a 0.42 decrease (from 0.80 to 0.38) in the absolute

SEM, but when the number of tasks was further increased from 5 to 12, there was a smaller decrease of 0.11 in the absolute SEMs. In contrast, the absolute SEM decreased only slightly when the number of ratings was increased from 1 to 2. The difference in the absolute SEM between the single- and double-rating situations varied from 0.05 to 0.14 for the section lengths of 1 to 12. The difference between the two became smaller as the section length increased. The same pattern of change was observed for the relative SEM [$\sigma(\delta)$], with the relative SEM being smaller than the absolute SEM for all the assessment scenarios, as expected. By definition, the relative SEM should usually be smaller than, or equal to, the absolute SEM. The SEMs from the $p \times$ T design [$\sigma(E)$] were slightly smaller than the relative SEM for the double-rating situations for various test lengths.

Figure 3 shows the 68% and 95% confidence intervals (CIs) based on the absolute SEMs [$\sigma(\Delta)$] for the universe score of 3 for single- and double-rating scenarios. The widths of the CIs for the double-rating scenario were somewhat narrower than those for the single-rating situation, but the difference between the two was not very large.

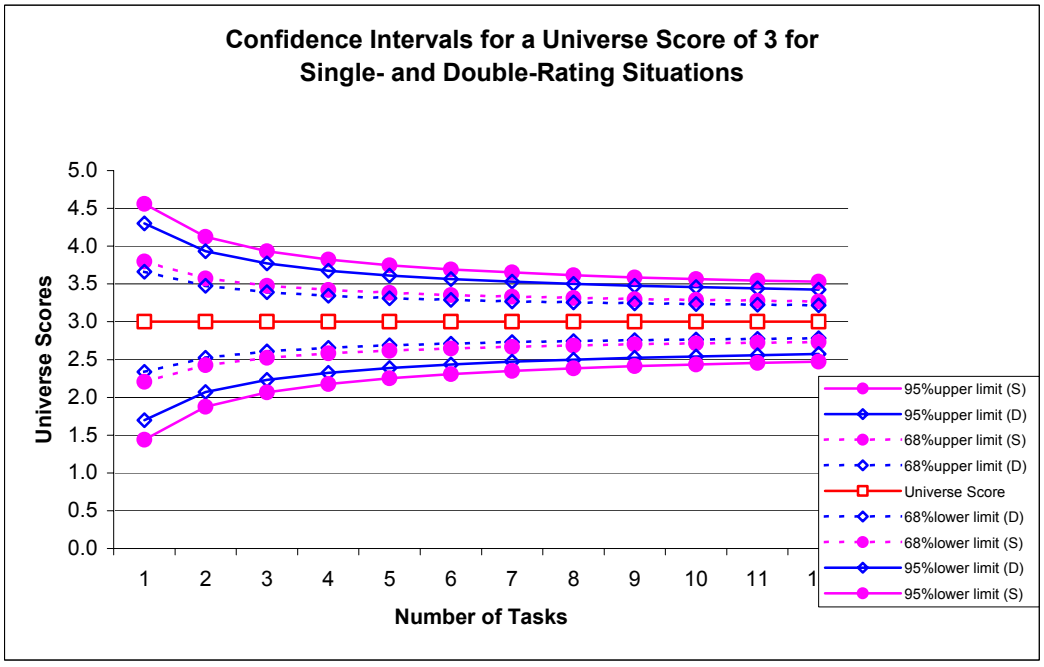


Figure 3. **Confidence intervals for a universe (true) speaking score of 3, based on absolute SEM [$\sigma(\Delta)$] for single- and double-rating situations from the univariate analysis.**

When the number of tasks was assumed to be 5, the 95% CI for the universe score of 3 was 2.25 ~ 3.75 (i.e., $3 \pm 1.96 \times 0.38$) for the single- rating scenario, but 2.39 ~ 3.61 for the double-rating scenario. It should be noted that the 95% CI for the 1-task-single-rated scenario was much wider (1.44 ~ 4.56) than those for the 5-tasks-and single-rated scenarios. When the number of tasks was increased to 12, the 95% CI for the same universe score of 3 shrank to 2.47 ~ 3.53 for the single-rating scenario and to 2.57~3.43 for the double-rating scenario. The 68% confidence intervals for the same universe score were somewhat narrower than the 95% confidence intervals, as expected. A similar pattern was observed for the confidence intervals based on the relative SEMs [$\sigma(\Delta)$] from the $p \times \mathrm{T} \times \mathrm{R}'$ design.

Focused comparison of two assessment scenarios. Table 4 shows the variance components and percentage of each variance component contributing to the total score variance, error variance, and reliability coefficients for two D-study scenarios of specific interest (i.e., 5-tasks-single-rated, 12-tasks-double-rated) in the $p \times \mathrm{T} \times \mathrm{R}'$ design.

Table 4

Focused Comparison of Two D-Study Assessment Scenarios From Original Data ($p \times T \times R'$)

	Assessment scenarios			
	5 tasks, single rating ($p \times \mathrm{T} \times \mathrm{R}'$)		12 tasks, double rating ($p \times \mathrm{T} \times \mathrm{R}'$)	
Effects	Variance	Percent	Variance	Percent
Examinee (p)	0.669	82.1	0.669	93.4
Task (T)	0.004	0.5	0.002	0.3
Rating (R′)	0.000	0.0	0.000	0.0
Examinee-by-task (pT)	0.045	5.5	0.019	2.6
Examinee-by-rating (pR′)	0.024	2.9	0.012	1.7
Task-by-rating (TR′)	0.001	0.1	0.000	0.0
Examinee-by-task-by-rating (pTR′, undifferentiated)	0.072	8.8	0.015	2.1
Total	0.814	100.0	0.716	100.0
Relative error (δ)	0.141	17.3	0.046	6.4
Absolute error (Δ)	0.146	17.9	0.048	6.6
G-coefficient ($E\rho^2$)	0.83		0.94	
Phi-index (Φ)	0.82		0.93	

As shown in Table 4, when the number of tasks was 5 for a single-rating situation, the percentage of the universe score (examinee) variance contributing to the total score variance was 82.1%, whereas those for the relative and absolute error variances were 17.3% and 17.9%, respectively. Given that the percentage of the examinee variance contributing to the total score variance in the G-study (for a single observation) was 51.3%, the difference between the two values can be seen as a significant increase considering the number of tasks was increased only from 1 to 5. When the numbers of both tasks and ratings were further increased from 5 to 12 and from 1 to 2, respectively, the percentage of the universe score (examinee) variance contributing to the total score variance increased further to 93.4%, while the percentage of the relative and absolute error variances decreased to 6.4% and 6.6%, respectively. Also shown in Table 4 are the estimates of the $E\rho^2$ and Φ coefficients for two assessment scenarios. It turned out that the higher $E\rho^2$ and Φ coefficients were obtained for the second assessment scenario (0.94, 0.93) rather than the first scenario (0.83, 0.82). As previously explained in relation to Figure 3, the width of the 95% CI for the 12-tasks-double-rating scenario (2.57 ~ 3.43) was narrower than that for the 5-tasks-single-rating scenario (2.25 ~ 3.75), indicating that the universe score of 3 would be more dependable in the former scenario. Nevertheless, it should be pointed out that each examinee would get a total of 24 ratings (i.e., 12 tasks × 2 ratings) in the test in the former scenario, but would receive a total of only 5 ratings (i.e., 5 tasks × 1 rating) in the latter scenario. Given such a huge increase in the total number of ratings per examinee, the increase in the percentage of the universe score variance (an 11.3% increase) and score reliability (a 0.11 increase) in the former scenario does not seem to be very large.

Multivariate Analysis ($p\bullet \times t^o \times r'\bullet$, $p\bullet \times t^o$)

Estimated variances and covariances. Tables 5 and 6 present the variance and covariance components for the three task-type subsections (i.e., listening-speaking, reading-speaking, and independent speaking) and the universe score correlations among the task-type subsections estimated from the $p\bullet \times t^o \times r'\bullet$ and $p\bullet \times t^o$ designs, respectively. As shown in Table 5, the examinee variance [$\sigma^2(p)$] was the largest variance component in each subsection in the $p\bullet \times t^o \times r'\bullet$ design, explaining about 56.3%, 55.7%, and 53.4% of the subsection total variances in the LS, RS, and IS subsections, respectively. The second largest variance component was that for the examinee-by-task-by-rating interaction plus undifferentiated error [$\sigma^2(ptr', \text{undifferentiated error})$] in each of the three subsections. The third largest variance component was that for the examinee-

by-task interaction variance [σ^2(pt)], followed by the examinee-by-rating interaction [σ^2(pr′)] in both the LS and IS subsections; but surprisingly, the σ^2(pr′) component was the third largest, followed by the σ^2(pt) in the RS subsection. However, the variance for the task main effect [σ^2(t)] was very small, accounting for only 1.4%, 0.4%, and 0.3% of each of the subsection total score variances for each subsection in the G-study. The variance component for the main effect for ratings [σ^2(r′)] explained less than 1% of the subsection score variances in all three subsections.

When the σ^2(pr′) component was compared across the three subsections, the component for the LS subsection was the smallest in its percentage (2.6%) in the total subsection score variance, whereas that for the RS subsection was the largest (11.8%). In fact, the comparatively large σ^2(pr′) component in the RS subsection is very important in explaining why the $p\bullet \times T^o \times R'\bullet$ and $p\bullet \times T^o$ designs yielded different results about composite score reliability (this will be discussed in more detail in the "Summary and Discussion" section).

As shown in Table 6, the σ^2(p) component was also the largest variance component for each of the three subsections in the $p\bullet \times t^o$ design, based on averaged ratings, explaining about 67.2%, 73.8%, and 67.9% of the total variances in the LS, RS, and IS subsections, respectively. The second largest variance component was that for the examinee-by-task interaction plus undifferentiated error [σ^2(pt, undifferentiated error)] in each of the three subsections, explaining 31.2%, 25.6%, and 31.8% of the total subsection score variance. The smallest variance component was that for the main effect for the tasks [σ^2(t)], which accounted for only 1.6%, 0.5%, and 0.3% of the subsection score variance.

Tables 5 and 6 also show the estimated universe score correlations among the subsections estimated from the $p\bullet \times t^o \times r'\bullet$ and the $p\bullet \times t^o$ designs, which might be used as a basis of decision about combining three subsection scores into a single composite score. The universe score correlations among the subsections were very high for all of the subsection pairs in both designs. In the $p\bullet \times t^o \times r'\bullet$ design, the universe score correlations between the LS and RS subsections, between the RS and IS subsections, and between the LS and IS subsections were 0.98, 0.95, and 0.89, respectively. In the $p\bullet \times t^o$ design, however, they were 0.92, 0.88, and 0.85, respectively, which were somewhat lower than those from the $p\bullet \times t^o \times r'\bullet$ design. This is due to the confounding of the rating facet and the object of measurement in this design (the reason for this will be explained in detail in the "Summary and Discussion" section).

Table 5

Estimated Variance and Covariance Components for the G-study in the Multivariate Analyses ($p\bullet \times t^{o} \times r'\bullet$) for the New TOEFL Speaking Section, Based on the Prototyping Data

Effects	G-study ($p\bullet \times t^{o} \times r'\bullet$)					
	LS		RS		IS	
	Vari/cov	Percent	Vari/cov	Percent	Vari/cov	Percent
Examinee (p)	**0.842**	56.3	*0.977*		*0.892*	
	0.789		**0.773**	55.7	*0.951*	
	0.623		0.637		**0.580**	53.4
Task (t)	**0.021**	1.4				
			0.005	0.4		
					0.003	0.3
Rating (r′)	**0.003**	0.2				
	-0.007		**0.009**	0.7		
	0.002		-0.004		**0.001**	0.1
Examinee-by-task (pt)	**0.209**	13.9				
			0.158	11.4		
					0.172	15.9
Examinee-by-rating (pr′)	**0.038**	2.6				
	0.001		**0.163**	11.8		
	-0.006		0.010		**0.090**	8.3
Task-by-rating (tr′)	**0.000**	0.0				
			0.002	0.1		
					0.000	0.0
Examinee-by-task-by rating, undifferentiated (ptr′, undifferentiated)	**0.382**	25.6				
			0.278	20.0		
					0.241	22.1
Total variance	1.496	100.0	1.388	100	1.087	100.0

Note. Bold-faced elements on the diagonal line in the second, fourth, and sixth columns are variances. Elements below the diagonal in these three columns are covariances. Elements above the diagonal (italicized) in these three columns are correlations.

Table 6

Estimated G-study Variance and Covariance Components for the G-study in the Multivariate Analysis ($p\bullet \times t^o$) for the New TOEFL Speaking Section, Based on the Prototyping Data ($n_p = 261$, $n_{t(LS)} = 4$, $n_{t(RS)} = 2$, $n_{t(IS)} = 5$)

Effects	G-study ($p\bullet \times t^o$)					
	LS		RS		IS	
	Vari/cov	Percent	Vari/cov	Percent	Vari/cov	Percent
Examinee (p)	**0.861**	67.2	*0.920*		*0.846*	
	0.789		**0.855**	73.8	*0.878*	
	0.621		0.642		**0.625**	67.9
Task (t)	**0.021**	1.6				
			0.006	0.5		
					0.003	0.3
Examinee-by-task, undifferentiated (pt, undifferentiated)	**0.400**	31.2				
			0.297	25.6		
					0.293	31.8
Total variance	1.282	100.0	1.157	100.0	0.921	100.0

Note. Bold-faced elements on the diagonal line in the second, fourth, and sixth columns are variances. Elements below the diagonal in these three columns are covariances. Elements above the diagonal (italicized) in these three columns are correlations.

Estimated subsection and composite score reliabilities. Figures 4 and 5 display the estimated reliability coefficients for different subsection lengths based on univariate analyses ($p \times \mathrm{T} \times \mathrm{R}'$, $p \times \mathrm{T}$) for each of the three subsections as part of multivariate analyses. Because examinees' task scores that were the averages of two raters' ratings on each task were used as units of analysis in the $p \times \mathrm{T}$ design, the reliability coefficients estimated in the $p \times \mathrm{T}$ design are actually comparable to those for the double-rating situation in the $p \times \mathrm{T} \times \mathrm{R}'$ design.

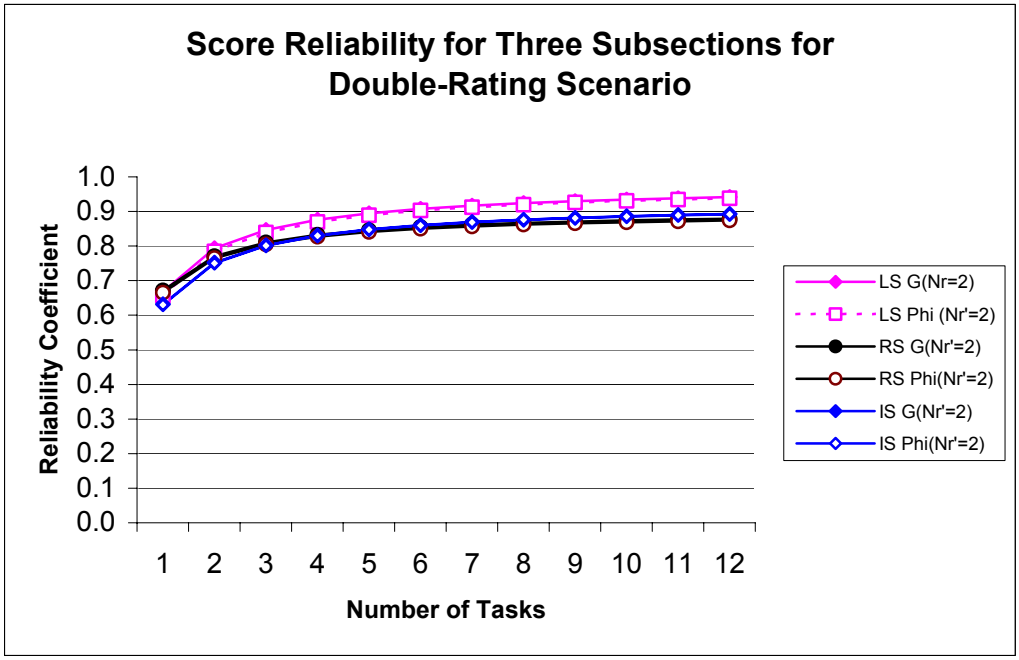


Figure 4. **Estimated reliability coefficients separately for each of the three subsections in multivariate analyses ($p\bullet \times T^{o} \times R'\bullet$) for different subsection lengths.**

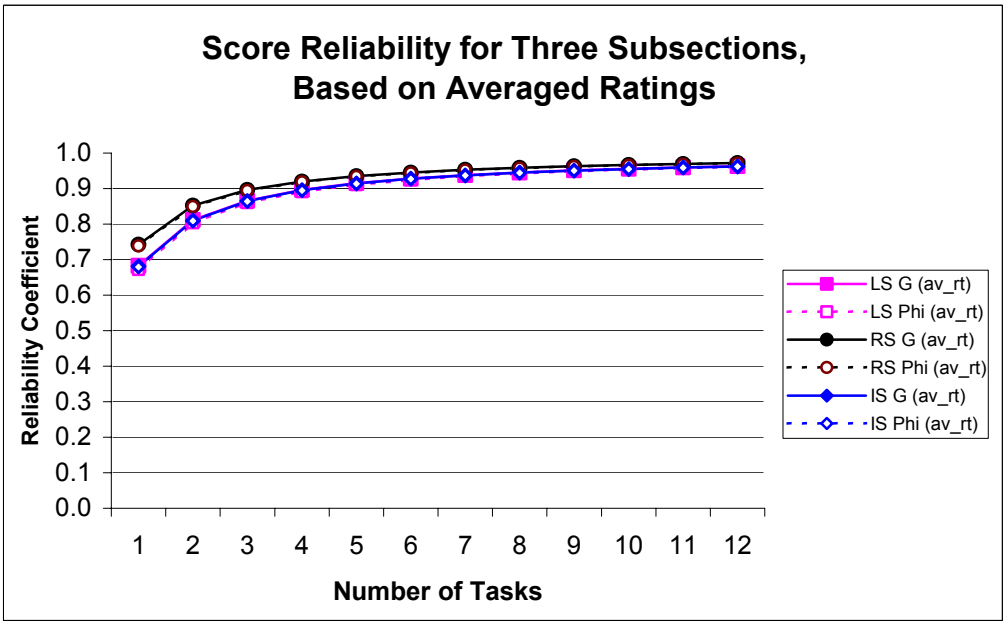


Figure 5. **Estimated reliability coefficients separately for each of the three subsections in multivariate analyses ($p\bullet \times T^{o}$) for different subsection lengths.**

One surprising result was that the $p \times T \times R'$ and $p \times T$ designs produced somewhat different reliability patterns among the three subsections in the speaking section. For instance, the LS subsection achieved a higher subsection score reliability than the RS and IS subsections for the same number of tasks for each subsection in the double-rating scenario, based on the $p \times T \times R'$ design, as shown in Figure 4. The same pattern was observed for the single-rating scenarios in the $p \times T \times R'$ design. In contrast, when the $p \times T$ design was used, the RS and LS sections produced higher score reliabilities than the IS subsection, as shown in Figure 5 (see the "Summary and Discussion" section for the explanation).

Table 7 and Figure 6 display the estimated reliability coefficients for the composite scores for various combinations of subsection lengths for fixed total section lengths in the single- and double-rating situations from the $p\bullet \times T^{o} \times R'\bullet$ design. Also shown in the same table and figure are the composite reliability coefficients for different section lengths estimated from the $p\bullet \times T^{o}$ design. The results indicate that the largest gains in composite score reliability would occur when the number of LS tasks was increased in the $p\bullet \times T^{o} \times R'\bullet$ design. Among the six scenarios for a fixed section length of 5 tasks, the scenario of 3 LS tasks, 1 RS task, and 1 IS task (3-1-1) achieved the highest $E\rho^2$ and Φ coefficients for both one- and two-rating situations. Similarly for the test length of 4 tasks, the highest $E\rho^2$ and Φ coefficients were obtained for the 2-1-1 scenario. However, the actual differences in score reliability values among different combinations of subsection lengths for the fixed section length of 5 tasks were not very large.

In the $p\bullet \times T^{o}$ design, however, the largest gain in composite score reliability was achieved when the number of RS tasks was increased. Among the six scenarios for the fixed section length of 5 tasks, the scenario of 1 LS task, 3 RS tasks, and 1 IS task (1-3-1) produced the highest $E\rho^2$ and Φ coefficients. Similarly for the test length of 4 tasks, the highest $E\rho^2$ and Φ coefficients were obtained for the 1-2-1 scenario. However, the actual differences in score reliability values among different combinations of subsection lengths for the fixed section length of 5 tasks were again rather small (see also the "Summary and Discussion" section).

Table 7

Estimated Generalizability Coefficients ($E\rho^2$) and Dependability Indices (Φ) for Composite Scores for Different Combinations of Subsection Lengths for Some Fixed Total Section Lengths

D-study scenarios				$p\bullet \times T^o \times R'\bullet$				$p\bullet \times T^o$	
				$n_{r'}=1$		$n_{r'}=2$		Averaged rating	
Number of tasks				$E\rho^2$	Φ	$E\rho^2$	Φ	$E\rho^2(\alpha_T)$	Φ
Total	LS	RS	IS						
3	1	1	1	0.78	0.78	0.85	0.84	0.87	0.86
4	2	1	1	0.83	0.82	0.88	0.88	0.90	0.89
4	1	2	1	0.81	0.81	0.87	0.87	0.90	0.90
4	1	1	2	0.81	0.81	0.87	0.87	0.89	0.89
5	3	1	1	0.85	0.85	0.90	0.90	0.91	0.91
5	2	2	1	0.84	0.84	0.90	0.89	0.92	0.92
5	2	1	2	0.85	0.84	0.89	0.89	0.91	0.91
5	1	2	2	0.83	0.83	0.89	0.89	0.92	0.92
5	1	1	3	0.83	0.83	0.88	0.88	0.91	0.91
5	1	3	1	0.82	0.82	0.88	0.88	0.92	0.92
6	2	1	3	0.86	0.85	0.90	0.90	0.93	0.92
6	1	1	4	0.84	0.84	0.89	0.89	0.93	0.92
11	4	2	5	0.90	0.90	0.94	0.94	0.96	0.96
12	4	3	5	0.91	0.90	0.94	0.94	0.96	0.96

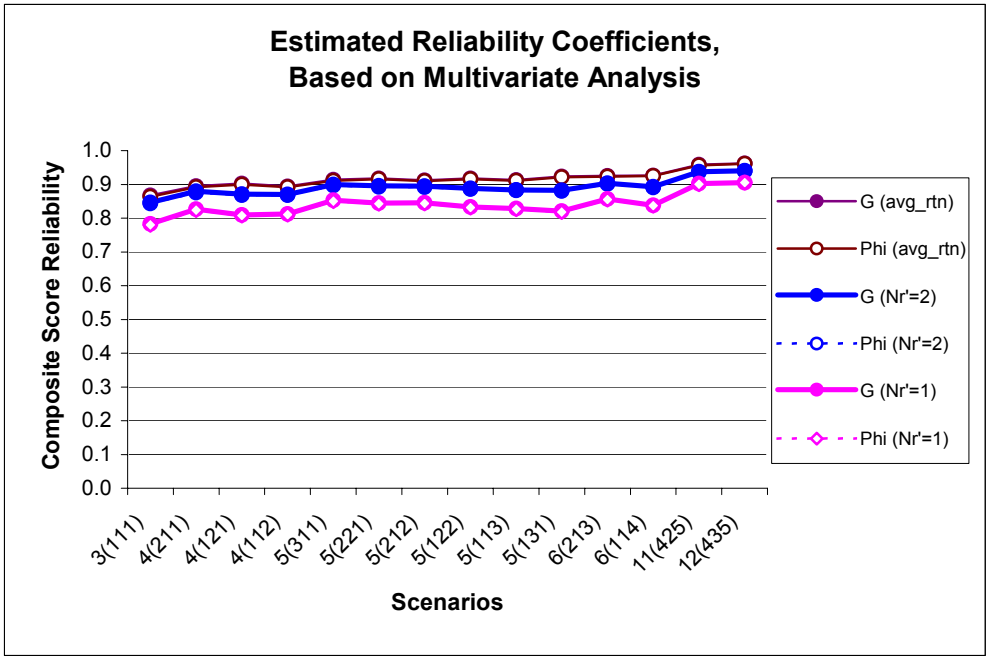


Figure 6. **Estimated reliability coefficients for section composite scores, based on multivariate analyses ($p\bullet \times T^{0} \times R'\bullet$; $p\bullet \times T^{0}$) for single- and double-rating scenarios for different combinations of subsection lengths for fixed total section lengths.**

Summary and Discussion

The purpose of the study was to examine (a) the relative effects of tasks and raters on examinees' speaking scores based on integrated and independent tasks, and (b) the impact of subsection lengths as well as the number of tasks and raters on the dependability of speaking scores in the G-theory framework. It was found that (a) the largest portion of error variance was related to tasks rather than raters in the study, (b) increasing the number of tasks had a relatively large impact on the score dependability up to a point of diminishing return, (c) the high universe score correlations among three subsections provided justification for combining the task-type subsection scores into a single composite score, and (d) slightly larger gains in composite score reliability were achieved when the number of LS (listening-speaking) tasks was increased, but the actual reliability differences among various combinations of subsection lengths for fixed total lengths were not large. These findings are discussed next in more detail.

Relative Effects of Examinees, Tasks, and Raters

Overall, expected patterns were observed regarding the relative effects of examinees, tasks, and raters on speaking scores in this study. Both univariate and multivariate analyses showed that the largest source of variation in examinees' test performances was attributable to differences among examinees in speaking ability measured by the test. In the univariate analyses based on the two-facet designs, the variance associated with the examinee main effect explained about a half (51.3 %) of the total section score variance in the G-studies. As the total number of speaking tasks increased up to 5 in the single-rating-per-sample scenario, the percentage of the examinee variance contributing to the total score variance increased drastically to 82.1%. This examinee variance component estimated in the G-study becomes a universe (or true) score variance later in the D-study. This means that, as intended, the tasks do distinguish among examinees on the construct measured by these tasks as a whole. A similar pattern also appeared in each task-type subsection in the multivariate analyses based on the two-facet design. When each of the subsections was examined separately, the examinee score variance explained a bit more than half of the subsection total variances (56.3%, 55.7%, and 53.4% in the LS, RS, and IS subsections, respectively). It should be noted that the percentage of the examinee variance contributing to the subsection total variance was the largest in the LS subsection, which suggests that the LS subsection is discriminating examinees slightly better than the RS (reading-speaking) and IS (independent speaking) subsections in this study.

In the univariate analysis based on the single-facet design ($p \times t$) on averaged ratings, the examinee variance estimated in the G-study occupied a considerably larger portion (61.3%) of the total score variance than in the two-facet design ($p \times t \times r'$). Even in the multivariate analyses based on the single-facet design, the examinee variance components explained significantly larger portions of the total subsection score variances (67.2%, 73.8%, 67.9%) than in the double-facet design. This was very much expected, because the basic units of analyses in the G-study for the single-facet design were examinees' task scores that were the averages of the two raters' ratings on each task. For this reason, it would be fairer to compare the proportions of the G-study examinee variance in the single-facet designs ($p \times t$, $p\bullet \times t^{o}$) with those of the D-study examinee variance for the single-task-and-double-rating scenario in the two-facet designs ($p \times T \times R'$, $p\bullet \times T^{o} \times R'\bullet$). Nevertheless, it was subsequently found that the proportions of the G-study

examinee variance in the single-facet designs were slightly larger than even those of the corresponding D-study examinee variance in the two-facet designs.

This may be partly due to the fact that the rating (r′) facet is treated as a hidden fixed facet in the p × t and p• × t^{o} designs and thus becomes a part of the object of measurement (Brennan, 2001; Suen, 1990). When this happens, the error variances attributable to the ratings are absorbed into the universe score (examinee) variance in the p × t and p• × t^{o} designs, making the proportion of the examinee variance contributing to the total variance become somewhat larger than it is for an equivalent D-study (double-rating) scenario in the $p \times T \times R'$ and $p\bullet \times T^{o} \times R'\bullet$ designs. For this reason, the universe score (examinee) variance in the p × t and p• × t^{o} designs becomes less meaningful due to the confounding of the examinee- and rating-related error variances.

The largest source of relative error variance in this study turned out to be tasks. In the univariate analyses based on the two-facet designs, the second largest variance component contributing to error variance was found to be the one associated with the examinee-by-task interaction, which accounted for about 17% of the total score variance. This means that different tasks used in this study are not rank-ordering the examinees consistently in the same order on the rating scale. Nonetheless, the variance associated with the main effect for tasks was very small in both the (r:p) × t (1.8%) and the p × t × r′ (1.7%) designs, suggesting that tasks might vary only slightly in difficulty and thus be regarded as comparable overall. In the multivariate analysis (p• × t^{o} × r′•), a similar pattern was also observed in each task-type subsection. The examinee-by-task interaction variance was the second largest error variance component for the LS and IS subsections (14%, 16%) and the third largest error variance component for the RS subsection (11%). Nonetheless, the variance component associated with the task main effect explained less than 1.4%, 0.4%, and 0.3% of the LS, RS, and IS subsection total variances, respectively, which suggests that tasks are similar in difficulty even within each subsection.

As a matter of fact, the largest source of error variance was the equivalent of the three-way interaction plus undifferentiated error variance [i.e., σ^{2}(tr:p, undifferentiated) or σ^{2}(ptr′, undifferentiated)] in both the univariate and multivariate analyses based on the two-facet designs. In the univariate analyses [(r:p) × t, p × t × r′], this variance component explained more than a quarter of the total variance in each design. In the multivariate analysis (p• × t^{o} × r′•), the

examinee-by-task-by-rating interaction plus undifferentiated error variance again accounted for more than a fifth of the subsection total score variances in each of the three subsections. However, it should be noted that this large variance component is related not only to tasks and raters but also to undifferentiated error.

The study found that raters (or ratings) did contribute to a total score variation in the speaking section to some extent, but their overall effects on the speaking scores were relatively small, compared to those of tasks. In the partially nested, univariate design [(r:p) × t], the main effect for raters (nested within examinees) [σ^2(r:p)] was the smallest variance component (explaining less than 2% of the total variance). As previously mentioned, due to the confounding of two different kinds of effects (i.e., the rater main effect and the examinee-by-rater interaction effect) in this variance component, it was not possible to tease out how much of this small rater-related effect can be ascribable to differences in severity among raters and rater inconsistency across examinees in the (r:p) × t design. A similar conclusion could be made based on the results of both univariate and multivariate analyses based on the two-facet crossed designs (p × t × r′, p• × t^{o} × r′•). The main effect for ratings was virtually zero in the p × t × r′ design, while it accounted for a very small proportion of the total subsection score variances in each subsection in the p• × t^{o} × r′• design (0.2%, 0.7%, and 0.1%, respectively, in the LS, RS, and IS subsections). This means that there was near zero difference in overall severity between the first and second ratings. This was not very surprising, because the same individual raters were allowed to serve as the first raters for some examinees and the second raters for other examinees in the same rating sessions. It is possible that some effects of potential severity differences among raters could have been accumulated and canceled out across examinees in a similar fashion in the first and second ratings. In addition, some of the remaining effects of the rater severity differences could have been captured into the examinee-by-rating interaction variance.

In relation to this, one intriguing finding was that the proportion of the examinee-by-rating interaction variance contributing to the total variance showed somewhat different patterns in the univariate and multivariate analyses. In the univariate analysis, the examinee-by-rating interaction variance explained only a small portion of the total score variance (less than 2%), which suggests that examinees were rank-ordered consistently across the first and second ratings in the test. In the multivariate analyses, however, the examinee-by-rating interaction variance was small in the LS subsection that was made up of 4 tasks (explaining about 3% of the total

variance), but rather substantial in the RS subsection consisting of only 2 tasks (about 12% of the total variance). This means that a significant portion of examinees were not rank-ordered consistently across the first and second ratings in the RS subsection, compared to the LS subsection. This can explain why the proportion of the universe score variance in the total subsection score variance in the G-study was the largest for the RS subsection in the $p\bullet \times t^{o}$ design, even though it was the largest for the LS in the $p\bullet \times t^{o} \times r'\bullet$ design. Since the ratings (r') are treated as a hidden fixed facet, the examinee-by-rating variance becomes absorbed into the examinee variance in the $p\bullet \times t^{o}$ design. For this reason, the unusually large examinee-by-rating interaction variance in the RS subsection becomes a part of the examinee variance, changing the rank-orderings of the three subsections in terms of the proportion of the universe score variance.

All these taken together, low score generalizability across tasks seemed to be the largest source of error variance in this study when all of the tasks in the test were examined together as a whole. Nevertheless, when each task-type subsection was examined separately, the rating inconsistency turned out to be an equally serious source of error variance along with the low task generalizability, particularly in the RS subsection. A more in-depth investigation might be necessary to examine whether such a distinct pattern is a result of the small number of tasks (2 tasks) sampled in the RS subsection or due to any systematic effect related to the interaction between raters and tasks in this subsection.

Impact of Number of Tasks and Raters on Score Dependability

This study found that increasing the number of tasks had a relatively large impact on the score dependability up to a point of diminishing return, but that the impact of one rating per speech sample on the score reliability seemed to be very small. In the univariate analysis, when the number of tasks was increased from 1 to 5 for the single-rating scenario, a dramatic increase (from 0.51 to 0.82) occurred in the dependability index. However, there seemed to be a diminishing return in score reliability for increasing the number of tasks beyond 5 tasks. When the number of tasks was further increased to 12, there was a much smaller increase (to 0.90). The expected pattern was also observed in the SEMs for increasing the number of tasks. The decrease in SEM was much larger when the number of tasks was increased from 1 to 5 than when the number of tasks was further increased to 12. In contrast, the dependability index was increased only slightly when the number of ratings was increased from 1 to 2. As the section length

became longer, this increased portion of reliability became smaller. Similarly, the decrease in the absolute SEM due to the double-rating scheme was also small for various section lengths.

Even when comparisons were made between the single- and double-rating testing scenarios, with the total number of ratings per examinee in the test held constant, increasing the number of tasks also turned out to have a larger impact on score reliability than the number of ratings per speech sample. Such a small impact of a double-rating scheme on score dependability was also ascertained in the confidence intervals based on the absolute SEMs for a particular universe score of 3 for different test lengths ranging from 1 to 12 tasks. The widths of the confidence intervals for the double-rating scenario were somewhat narrower than those for the single-rating situation, but the difference between the two was rather small.

In addition, a focused comparison of two particular assessment scenarios revealed that the same universe scores would be more dependable in the 12-tasks-and-double-rating scenarios than the 5-tasks-and-single-rating scenario. Higher generalizability coefficients and dependability indices were obtained for the first assessment scenario (0.94, 0.93) than for the second scenario (0.83, 0.82). The width of the 95% CI for the 12-tasks-and-double-rating scenario was narrower than that for the 5-tasks-and-single-rating scenario, as expected. Given a huge increase in the total number of ratings per examinee in the test for the first scenario (i.e., from 5 ratings to 24 ratings), however, the increase in the score dependability was not very large for this particular assessment scenario.

Overall, the results of the current study are consistent with the findings of previous research in performance-based assessments in general (Gao, Shavelson, & Baxter, 1994; Linn, 1993; Linn, Burton, DeStefano, & Hanson, 1996; Miller & Linn, 2000; Shavelson, Baxter, & Gao, 1993) and performance-based writing (Breland et al., 1999; Brennan et al., 1995; Dunbar, Koretz, & Hoover, 1991) and speaking assessments in particular (Fulcher, 2003; Lee et al., 2001). In most of the previous research based on holistic scoring of performance samples, rater variance components and their interaction components were also found to be relatively small compared to the examinee-by-task interactions, resulting in fewer raters needed to achieve acceptable values of score reliability or generalizability in large-scale, performance-based assessment than might be expected.[7]

Justifiability of Combining Subsection Scores Into a Single Composite Score

The universe scores from the three subsections estimated based on both the $p\bullet \times t^{o} \times r'\bullet$ and $p\bullet \times t^{o}$ designs were very highly correlated, providing a good justification for combining the subsection scores into a single composite score. The universe score correlation between the LS and RS subsections was close to a perfect correlation (0.98) in the first design and very high (0.92) in the second design, as well. Such very high correlations between the two subsections of integrated task types suggest that both LS and RS task types are, in essence, measuring a very similar (or the same) underlying construct (i.e., speaking proficiency), even though the input stimuli modes are different (auditory texts vs. reading passages).

The universe score correlations between the RS and the IS subsections were also very high (0.95 and 0.89 in the first and second G-study designs), even though it was assumed that these two task types were theoretically intended to tap somewhat different ranges of speaking subskills. The only commonality between the two task types was that the input stimuli were of the visual nature (reading passages vs. visuals, such as pictures and maps). In contrast, the lowest universe score correlation was obtained between the LS and IS subsections (0.89 and 0.85 in both the first and second designs, respectively). It may be argued that such a comparatively lower universe score correlation between the LS and IS subsections might be due to the fact that the two task types are comparatively more dissimilar than other pairs in terms of both stimuli input modes (auditory vs. visual) and the nature of tasks (integrated vs. independent). Nevertheless, it should be pointed out that this comparatively lower correlation between the LS and IS subsections can still be regarded as being quite high.

As a matter of course, the universe score correlations estimated based on the $p\bullet \times t^{o}$ design were comparatively lower than those from the $p\bullet \times t^{o} \times r'\bullet$ design. As previously mentioned, the rating facet was treated as a hidden fixed facet and thus became a part of the object of measurement in the $p\bullet \times t^{o}$ design. When this happens, the proportion of the examinee variance contributing to the total subsection variance is larger than it is in $p\bullet \times t^{o} \times r'\bullet$ design, resulting in the universe score variance being less meaningful and more restricted in terms of generalizability. Because the subsection universe score variances are confounded with the error variances associated with the rating facet, it might be expected that the universe score correlations among the subsections would be somewhat lower in the $p\bullet \times t^{o}$ design than in the

p• × t^{o} × r′• design. In that sense, the estimates of universe correlations from the p• × t^{o} × r′• design should be more accurate and meaningful than those from the p• × t^{o} design.

Optimal Combinations of Subsection Lengths

In the multivariate analyses, the *p*• × T^{o} and *p*• × T^{o} × R′• designs produced slightly different patterns of composite score reliability for various combinations of subsection lengths for fixed total section lengths. In the *p*• × T^{o} × R′• design, it was found that the largest gains in composite score reliability occurred when the number of LS tasks was increased, largely because the examinee-by-rating interaction variance (rating inconsistency) was very small in the LS subsection. Among the six scenarios for the section length of 5 tasks, the scenario of 3 LS tasks and 1 RS task and 1 IS task (3-1-1) produced the highest dependability indices for both the single- (0.85) and double-rating situations (0.90). Similarly for the section length of 4 tasks, the highest dependability indices were obtained for the 2-1-1 scenario. This is consistent with the fact that the proportion of the universe score variance was the largest in the LS subsection, with the proportion of the relative and absolute error variance being the smallest. Moreover, the LS subsection achieved the highest subsection score reliability among the three subsections, given the same subsection length. This is partially due to the fact that the examinee-by-rating interaction variance component in the LS subsection was smaller in size and proportion than in other two subsections, while the relative proportions of other error variance components were similar across the three subsections. However, the actual differences in score reliability values among different combinations of subsection lengths for a fixed section length were not significantly large.

As noted earlier, the *p*• × T^{o} design produced slightly different results from the *p*• × T^{o} × R′• design. The largest gains in composite score reliability were achieved when the number of RS tasks was increased in the *p*• × T^{o} design. It should be mentioned, however, that the estimates of composite score reliabilities from the *p*• × T^{o} × R′• design might be more accurate than those from the *p*• × T^{o} design, as the rating facet is properly taken into account as a random facet in modeling measurement error in the former design.

Conclusions and Avenues for Future Research

Conclusion

Univariate analyses have shown that, to maximize score reliability for speaking, it would be more efficient to increase the number of tasks than the number of ratings per speech sample. The tasks in this study do distinguish among examinees in terms of the speaking construct to be measured by the test. While the tasks are, on average, comparable in difficulty, they are not uniformly difficult for all examinees. The difference in rater severity between the first and second ratings was negligible, and examinees were rank-ordered in a similar way across the first and second ratings overall. As a result, adopting a single-rating scheme had a relatively small effect on the score dependability. Moreover, it seems that the reduced portion of score reliability resulting from the adoption of the single-rating scheme could probably be compensated for by increasing the number of tasks. Clearly, however, increasing the number of tasks beyond 5 or 6 tasks would result in diminishing returns.

Multivariate G-theory analysis has provided very useful, additional information about the justifiability of reporting a composite score for the whole section and the optimal configurations of the speaking tasks. First, given the high universe score correlations among the three subsections in both designs, it seems that it is justifiable to combine these subsection scores into a single composite speaking score from the score dependability point of view. Second, more gains in composite score reliability could result from increasing the number of listening-speaking tasks for the fixed section lengths. However, the actual differences in score reliability values among different combinations of the number of tasks in each subsection for the fixed total section length were not large. Therefore, the final decisions about the possible test configurations should also be based on other factors, such as content and item-development considerations.

Methodologically, the multivariate analysis has also demonstrated the importance of properly modeling the rater and task effects simultaneously in the investigations of score reliability and measurement error in rater-mediated language assessment. When the rater (rating) effect was not appropriately taken into account in the $p\bullet \times t^{o}$ design based on averaged ratings, it actually produced quite different results from the $p\bullet \times t^{o} \times r'\bullet$ design, as shown in this study. Because the score variability attributable to the ratings was absorbed into the universe score

variance, somewhat distorted results were obtained for the universe score correlations and the composite score reliability in the $p\bullet \times t^{o}$ design.

Avenues for Further Investigation

Replication with rerated speech samples. As previously mentioned, a fully crossed design with tasks and raters as random facets ($p \times t \times r$) is advantageous in the G-study, because it can maximize the number of design structures that can be considered in the D-study (Brennan, 2001). In this study, however, a two-facet, partially nested design [(r:p) × t] was used along with a two-facet crossed design ($p \times t \times r'$) in the univariate analysis. It was demonstrated that the two designs produced almost identical results in terms of measurement error and score reliability. In both designs, however, the data did not allow for separate estimation of the main effect for raters (rater severity) and the person-by-rater interaction (rater inconsistency) effect. For this reason, it was not really possible to investigate the impact on the score dependability of using different nested rating scenarios (e.g., using a single rating per speech sample but having each task for the particular test taker rated by a different rater) through the two designs. To conduct such an analysis, it would be necessary to obtain a complete data matrix by having the responses of a sample of examinees rerated by multiple raters according to the $p \times t \times r$ design.

Replication with a larger, balanced number of tasks in each subsection. The power of generalizability can also be realized when there is a large sample of observations available for each facet in the universe of admissible observations. In this study, only 2 tasks were included in the RS subsection. It was found that the proportion of the examinee-by-rating interaction variance was larger in the RS subsection than in the other two subsections. One interesting question is whether such a distinct pattern is due to the small number of tasks (only 2 tasks) sampled in the RS subsection or to any real systematic effect related to the raters and task types. If a larger number of tasks had been sampled for the RS subsection, the variance estimates for the subsection would have been more stable and thus would have been able to strengthen the generalizability of the results about the rater-related error. In addition, if the same number of tasks were included in all of the three subsections, as in a balanced design, a fairer comparison might have been possible among the three subsections.

References

Bachman, L. F. (1990). *Fundamental considerations in language testing.* New York: Oxford University Press.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgment in a performance test of foreign speaking. *Language Testing, 12*, 238-257.

Bolus, R. E., Hinofotis, F. B., & Bailey, K. M. (1982). An introduction of generalizability theory in second language acquisition research. *Language Learning*, *32*(2), 245-258.

Breland, H., Bridgeman, B., & Fowles, M. E. (1999). *Writing assessment in admission to higher education: Review and framework* (ETS RR-99-3). Princeton, NJ: ETS.

Brennan, R. L. (1992). *Elements of generalizability theory.* Iowa City, IA: ACT.

Brennan, R. L. (1999). mGENOVA (version 2.0) [Computer software]. Iowa City, IA: The University of Iowa, Iowa Testing Programs.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24*(4), 339-353.

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

Brennan, R. L., Gao, X., & Colton, D. A. (1995). Generalizability analyses of Work Keys listening and writing tests. *Educational and Psychological Measurement*, *55*, 157-176.

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL® 2000 speaking framework: A working paper* (TOEFL Monograph No. MS-20). Princeton, NJ: ETS.

Crick, G. E., & Brennan, R. L. (1983). *GENOVA* [Computer software]. Iowa City, IA: The University of Iowa, Iowa Testing Programs.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability.* New York: John Wiley.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL® 2000 writing framework: A working paper* (TOEFL® Monograph No. MS–18; ETS RM-00-05). Princeton, NJ: ETS.

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,* 298-303.

Enright, M. K., Bridgeman, B., Cline, F., Eignor, D., Lee, Y., & Powers, D. (2003). *Evaluating measures of communicative language ability.* Unpublished manuscript. Princeton, NJ: ETS.

Fulcher, G. (2003). *Testing second language speaking.* Essex, England: Pearson Professional Education.

Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education, 7*(*4*), 323-342.

Henning, G. (1996). Accounting for nonsystematic error in performance testing. *Language Testing, 13*, 53-61.

Lee, Y.-W., Golub-Smith, M., Payton, C., & Carey, J. (2001, April). *The score reliability of the Test of Spoken English™ (TSE®) from the generalizability theory perspective: Validating the current procedure.* Paper presented at the annual conference of American Educational Research Association (AERA), Seattle, WA.

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

Lewkowicz, J. A. (1997). The integrated testing of a second language. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education: Vol. 7. Language testing and assessment* (pp. 121-130). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Linn, R. L. (1993). Performance-based assessments: Implications of task specificity. *Educational Measurement: Issues and Practice*, *13*(*1*), 5-8, 15.

Linn, R. L., Burton, E., DeStefano, L., & Hanson, M. (1996). Generalizability of new standards project 1993 pilot study tasks in mathematics. *Applied Measurement in Education, 9*(*3*), 201-214.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*(2), 158-80.

Miller, M. D., & Linn, R. L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement, 24*, 367-378.

Read, J. (1990). Providing relevant content in an EAP writing test. *English for Specific Purposes, 9*, 109-121.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance-based assessments. *Journal of Educational Measurement, 30*, 215-232.

Shavelson, R. J., & Webb, N. R. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Suen, H. K. (1990). *Principles of test theories.* Hillsdale, NJ: Lawrence Erlbaum.

van Weeren, J., & Theunissen, T. J. J. (1987). Testing pronunciation: An application of generalizability theory. *Language Learning, 37*, 109-122.

Weir, C. J. (1993). *Understanding and developing language tests.* Hemel Hamstead, England: Prentice Hall.

Wesche, B. (1987). Second language performance testing: The Ontario test of ESL as an example. *Language Testing, 4,* 28-47.

Notes

[1] Many-Faceted Rasch Measurement (MFRM) can be viewed as an alternative to generalizability theory analyses in examining the effects of tasks and raters on examinees' scores. However, the focus of this research is to investigate the impact of such facets on score dependability for various assessment scenarios through D-studies. The MFRM approach, while providing more detailed diagnostic information at the levels of individual examinees, tasks, raters, and combinations of these elements, does not lend itself well for such investigation as extrapolating to assessment scenarios that are different from the one used in the data-collection process.

[2] Often, a statistical procedure of *test equating* is conducted for multiple-choice sections of a large-scale language test (e.g., TOEFL) to make an adjustment for form-to-form difficulty differences. Such a procedure makes it possible for test scores to have equivalent meanings across test forms. Under such circumstance, a generalizability coefficient can represent measurement accuracy for both norm-referenced and criterion-referenced score interpretations. However, because performance-based testing involves only a small number of tasks and subjective rater judgment in scoring, test equating is often not a feasible option for performance-based speaking assessment. In addition, the dependability index is a rather conservative (safer) estimate of score reliability, compared to the generalizability coefficient. For this reason, a dependability index is often a preferred type of reliability coefficient for rater-mediated performance assessments.

[3] In generalizability theory notation, measurement facets in a generalizability study (G-study) are identified by lowercase letters (e.g., t and r), but the facets in a decision study (D-study) are identified by uppercase letters (e.g., T and R). However, the examinee (p) facet, which is the object of measurement in this study, is represented by a lowercase letter in both G- and D-studies. It should be noted, however, that the italicized letter (*p*) is usually used for the object of measurement in the D-study. In addition, operator × means "crossed with," while : means "nested within" in both G- and D-studies. In the p × t × r design, for instance, persons are crossed with tasks that are crossed with raters; whereas raters are nested within persons in the (r:p) × t design, even though tasks are crossed with both persons and raters.

[4] In the multivariate design, a superscript filled-in circle (•) next to a facet symbol indicates that the facet is crossed with the fixed-category facet (v), whereas a superscript empty circle (o) signals that the facet is nested within the multivariate variables (v), which is a task-type subsection facet in this study.

[5] The 5-tasks-rated-once scenario was rather arbitrarily chosen in this study, because this was one of the single-rating assessment scenarios most favored by test development staff in terms of satisfying task and test design constraints at the beginning of the study. Other equally favored section-length scenarios were section lengths of 3, 4, and 6 tasks that were examined in the multivariate analyses in more detail. In contrast, the 12-tasks-rated-twice scenario was selected because this may represent the assessment scenario for the current tape-mediated TSE.

[6] If we look at the p × t design from the perspective of the p × t × r′ design, ratings (r′) should be treated as a hidden fixed facet in the p × t design. In the p × t × r′ design, ratings (r′) are assumed to be crossed with examinees (p). This means that there would be the same two ratings (i.e., first ratings, second ratings) for all the examinees, at least in the formal representation of the rating facet. However, it should be noted that when we look at the p × t design from the (r:p) × t design perspective, what is hidden in the p × t design is actually the rater (r) facet, not the rating (r′) facet. Since there are different sets of raters for different examinees, raters (r) should be treated as a hidden random facet in this perspective.

[7] As one reviewer has pointed out, some of the previous research in rater-mediated language assessment has reported the existence of significant rater-related effects on examinees' speaking scores (Bachman et al., 1995; Bolus et al., 1982; Lynch & McNamara, 1998). It should be emphasized that the small rater-related effect confirmed in this study is a relative concept (i.e., when score variability associated with raters is proportionally compared to that associated with holistically scored tasks). In this sense, some of these previous studies are limited in the sense that (a) only raters and occasions were compared as random facets (Bolus et. al., 1982); (b) the two locally dependent tasks scored only on the grammar dimension were compared with raters as random facets (Bachman et al., 1995); and (c) analytic-dimension-by-task combinations (with built-in double-dependency structures among the combinations) were

compared with raters as random facets (Lynch & McNamara, 1998). When the focus of evaluation in rating speech samples is on the language-related dimensions of examinee performance (e.g., pronunciation, vocabulary, grammar), as it was in some of these studies, it may be very likely that examinee performance is consistent across different tasks and that scores are relatively more generalizable across tasks. It should be noted, however, that task generalizability can become a real issue when the "content/topic development" or "task-fulfillment" aspects of examinee performance is the important rating criterion along with other dimensions, as in the holistic scoring of examinee performance samples for performance-based assessment.

List of Appendixes

Page

Appendix A

Mathematical Formula for Computing Generalizability Coefficients ($E\rho^2$) and Dependability Indices (Φ) From the Univariate Analyses in the Study

In the G-study, the variances associated with various facets of measurement, including the object of measurement, are estimated and evaluated in terms of their relative importance in the total score variance. Three different designs were used to analyze the new TOEFL prototyping study: (r:p) × t, p × t × r′, and p × t. There could be a total of five, seven, and three variance components for each of the three designs, respectively, as follows:

1. (r:p) × t design: σ^2(p), σ^2(t), σ^2(r:p), σ^2(pt), σ^2(tr:p, undifferentiated)
2. p × t × r′ design: σ^2(p), σ^2(t), σ^2(r′), σ^2(pt), σ^2(pr′), σ^2(tr′), σ^2(ptr′, undifferentiated)
3. p × t design: σ^2(p), σ^2(t), σ^2(pt, undifferentiated)

In the D-study, two different kinds of score reliability equivalents can be computed for different measurement scenarios; that is, a generalizability coefficient ($E\rho^2$ or G) and a dependability index (Φ).

First, the relative error variance [$\sigma^2(\delta)$] and the generalizability coefficient ($E\rho^2$) can be defined for each of the designs, as in Equations 1a, 1b, and 1c, which can be interpreted as the error variance and the reliability coefficient for norm-referenced score interpretation, respectively (Brennan, 1992; Suen, 1990). In a single-facet design (p × t), a Cronbach alpha (α_T) is numerically equivalent to a $E\rho^2$ coefficient (Brennan, 1992; Suen, 1990).

(r:p) × t design:

$$\mathrm{E}\rho^2 = \frac{\sigma^2(p)}{\sigma^2(p)+\sigma^2(\delta)}$$

$$= \frac{\sigma^2(p)}{\sigma^2(p)+\frac{\sigma^2(r:p)}{n_r}+\frac{\sigma^2(pt)}{n_t}+\frac{\sigma^2(tr:p,undifferentiated)}{n_r n_t}} \qquad (1a)$$

p × t × r′ design:

$$\mathrm{E}\rho^2 = \frac{\sigma^2(p)}{\sigma^2(p)+\sigma^2(\delta)}$$

$$= \frac{\sigma^2(p)}{\sigma^2(p)+\frac{\sigma^2(pt)}{n_t}+\frac{\sigma^2(pr')}{n_{r'}}+\frac{\sigma^2(ptr',undifferentiated)}{n_t n_{r'}}} \quad (1b)$$

(p × t) design:

$$\mathrm{E}\rho^{2}{}_{(or\,\alpha_T)} = \frac{\sigma^2(p)}{\sigma^2(p)+\sigma^2(\delta)} = \frac{\sigma^2(p)}{\sigma^2(p)+\frac{\sigma^2(pt,undifferentiated)}{n_t}} \quad (1c)$$

Second, the absolute error variance [$\sigma^2(\Delta)$] and the dependability index (Φ) can be computed for the first two designs, as in Equations 2a and 2b, which can be interpreted as the error variance and the score reliability coefficient for criterion-referenced score interpretation, respectively. When the scores are given absolute interpretations, as in domain-referenced or criterion-referenced situations, the Φ coefficient is more appropriate (Brennan, 2001).

(r:p) × t design:

$$\Phi = \frac{\sigma^2(p)}{\sigma^2(p)+\sigma^2(\Delta)} =$$

$$= \frac{\sigma^2(p)}{\sigma^2(p)+\frac{\sigma^2(t)}{n_t}+\frac{\sigma^2(r:p)}{n_r}+\frac{\sigma^2(pt)}{n_t}+\frac{\sigma^2(tr:p,undifferentiated)}{n_r n_t}} \quad (2a)$$

p × t × r′ design:

$$\Phi = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\Delta)}$$

$$= \frac{\sigma^2(p)}{\sigma^2(p) + \frac{\sigma^2(t)}{n_t} + \frac{\sigma^2(r')}{n_{r'}} + \frac{\sigma^2(pt)}{n_t} + \frac{\sigma^2(pr')}{n_{r'}} + + \frac{\sigma^2(tr')}{n_t n_{r'}} + \frac{\sigma^2(ptr', undifferentiated)}{n_t n_{r'}}} \quad (2b)$$

Appendix B

Mathematical Formula for Computing Generalizability Coefficients ($E\rho^2$) and Dependability Indices (Φ) From the Multivariate Analyses in the Study

In the multivariate G-theory design (Brennan, 2001), different test subsections (*v, v′*) are viewed as different levels (or conditions) of a fixed facet, and the number of tasks in each level of the fixed facet can be the same (balanced) or different (unbalanced). It would be possible to estimate a set of variance components for each subsection separately and compute the $E\rho^2$ and Φ coefficients of the composite scores in the framework of multivariate G-theory. In addition, the covariance components can be computed for the facets that are crossed with the fixed-content subsection facet. In the context of the new TOEFL speaking section, an attractive option is to recognize the task-type subsections (e.g., LS, RS, IS) as a fixed facet in the multivariate $p\bullet \times t^o \times r'\bullet$ design.

Tables B1 and B2 show the variance and covariance components to be estimated for each subsection in both the $p\bullet \times t^o$ and the $p\bullet \times t^o \times r'\bullet$ designs.

Table B1

Variance and Covariance Components in the $p\bullet \times t^o$ Design

Effects	Variance and covariance components	
Examinee (*p*)	$\sigma_v^2(p)$	
	$\sigma_{vv'}(p)$	$\sigma_{v'}^2(p)$
Task (t)	$\sigma_v^2(t)$	
		$\sigma_{v'}^2(t)$
Examinee-by-task (pt, undifferentiated)	$\sigma_v^2(pt, undifferentiated)$	
		$\sigma_{v'}^2(pt, undifferentiated)$

Table B2

Variance and Covariance Components in the p• × t° × r′• Design

Effects	Variance and covariance components	
Examinee (*p*)	$\sigma_v^2(p)$	
	$\sigma_{vv'}(p)$	$\sigma_{v'}^2(p)$
Task (t)	$\sigma_v^2(t)$	
		$\sigma_{v'}^2(t)$
Rating (r′)	$\sigma_v^2(r')$	
	$\sigma_{vv'}(r')$	$\sigma_{v'}^2(r')$
Examinee-by-task (pt)	$\sigma_v^2(pt)$	
		$\sigma_{v'}^2(pt)$
Examinee-by-rating (pr′)	$\sigma_v^2(pr')$	
	$\sigma_{vv'}(pr')$	$\sigma_{v'}^2(pr')$
Task-by-rating (tr′)	$\sigma_v^2(tr')$	
		$\sigma_{v'}^2(tr')$
Examinee-by-task-by-rating (ptr′, undifferentiated)	$\sigma_v^2(ptr', undifferentiated)$	
		$\sigma_{v'}^2(ptr', undifferentiated)$

It should be noted that the fact that there are only two columns (v, v') does not necessarily mean that there are only two levels of the fixed facet. This compact form of notation is often used to represent the n_v levels of the fixed fact. In the context of the new TOEFL assessment, for instance, the variance and covariance components are estimated for each of the three levels for the fixed-content category facet (v^{LS}, v^{RS}, v^{IS}).

First, the relative error term for the composite score [$\sigma_C^2(\delta)$] and the composite score generalizability coefficient ($E\rho^2$ or G) can be defined, as in Equation 3, which can be interpreted as the error variance and the reliability index for norm-referenced score interpretation, respectively (Brennan, 1992).

$$\mathrm{E}\rho^2 = \frac{\sigma_C^2(\tau)}{\sigma_C^2(\tau) + \sigma_C^2(\delta)} = \frac{\sum_v \sum_{v'} \omega_v \omega_{v'} \sigma_{vv'}(\tau)}{\sum_v \sum_{v'} \omega_v \omega_{v'} [\sigma_{vv'}(\tau) + \sigma_{vv'}(\delta)]} \tag{3}$$

Second, the absolute error for the composite score [$\sigma_C^2(\Delta)$] and the composite score dependability index (Φ) can be computed, as in Equation 4, which can be interpreted as the error variance and the score reliability index for criterion-referenced score interpretation, respectively.

$$\Phi = \frac{\sigma_C^2(\tau)}{\sigma_C^2(\tau) + \sigma_C^2(\Delta)} = \frac{\sum_v \sum_{v'} \omega_v \omega_{v'} \sigma_{vv'}(\tau)}{\sum_v \sum_{v'} \omega_v \omega_{v'} [\sigma_{vv'}(\tau) + \sigma_{vv'}(\Delta)]} \tag{4}$$

In the new TOEFL speaking section, for instance, several different combinations of subsection lengths would be possible for a total section length of five tasks for a speaking section.

Appendix C
Sample Tasks for Integrated and Independent Task Types

Sample Listening-Speaking Tasks

Groundwater in the San Joaquin Valley

(N) Listen to part of a talk in an environmental science class. The professor is discussing groundwater.

Screen: Female professor

(Professor) We've been talking a lot about the importance of groundwater … as a … a critical natural resource. You may not realize that a large percentage of the water we use comes from underground. The amount of water stored in the rocks and sediments below the Earth's surface is … well, it's vast. And it's this water, groundwater, that's the primary source of … of everyday water usage around the world. In the United States alone, about half of our drinking water comes from underground … and it's at least 40% of the water that's used to irrigate farms. And it's a large part of the water used by various industries too. But … and this is the direction I want to take today … overuse of groundwater has created some serious environmental concerns … not just about its depletion … or even contamination of the water, but also the damage it causes to the land … to the surface of the land.

This damage to the land surface is generally referred to as *land subsidence*.

Screen: Blackboard: Land Subsidence

It's where large portions of the ground actually sink. Now sometimes, and your book talks about this too, sometimes *natural* processes related to groundwater can cause the ground to sink. But it also happens when too much water is pumped from underground wells … pumped faster than it can naturally be replaced. Now this is particularly true in areas with thick underground layers of loose sediment. What happens is … as the water gets pumped out, the water pressure drops. When this happens, the weight of the land is transferred to the sediment. And as the weight—you know, the pressure—on the sediment increases, the grains get packed more and more tightly together and this eventually causes the ground to sink.

Now a classic example of this happened in the San Joaquin Valley in California.

Screen: Blackboard: San Joaquin Valley

The San Joaquin Valley is an agricultural area—it makes up a large part of central California … and it's an area where large amounts of water are used to irrigate the crops. For a long time, almost half of the water used for irrigation there came from underground. And … in nearly every city in the area, groundwater was the main source of water—for industry and for the general population. What happened was … they started pumping groundwater in the late 1800s … and over time the amount of water they pumped gradually increased. So in about 20 or 30 years … by the 1920s, some land subsidence had already started. The pumping continued and by the early '70s, the amount of water being pumped had increased so much that water levels, underground water levels, had dropped nearly 120 meters … and the ground had sunk more than 8-and-a-half meters. Of course, this 8-and-a-half-meter drop was gradual over time but imagine a large area of land sinking so much. So finally … in the '70s, they decided to try some things to stop the problem. They tried … they reduced the amount of water they were pumping from underground but they needed another source of water so what they … they started importing water—they brought in *surface* water so for a few years, groundwater pumping slowed down quite a bit. And eventually, the water levels started to recover and the land subsidence seemed to stop. But, unfortunately, problems started again when the area was hit with a *drought* … that was in the '90s—not long ago. They were without rain for long periods of time, so the surface water they'd been relying on wasn't available. So what did they do? Well, they were forced to start pumping more groundwater again—which caused, again, more land subsidence. And this time, water levels dropped much faster … that nearly half of the entire valley was affected by land subsidence.

Another good example was in Mexico City, which is, of course, a heavily populated city. Thousands of wells were pumping water from underneath the city. And as more and more water was removed, parts of the city subsided by as much as 6 or 7 meters. Some buildings sank so much that now when you enter a building from the street … that used to be the second floor level.

Speaking Question

The professor describes a series of events that occurred in the San Joaquin Valley in California. Explain what happened there. In your response, be sure to include details about:

- the problems that occurred there
- the causes of the problems
- the efforts that were made to solve the problems

Sample Reading-Speaking Tasks

Innate Versus Learned Perception

The controversy over the importance of innate factors, or factors present at birth, versus the importance of learning in perceptual development led a number of psychologists to make a thorough examination of perception in the first few months of life. The prevalent view used to be that the perceptual experience of the infant was, in the words of philosopher and psychologist William James, a "booming, buzzing confusion." It is now generally acknowledged that infants possess greater perceptual skills than previously thought. Research carried out by Robert Fantz was important in changing the climate of opinion. He made use of a preference task in which a number of visual stimuli were presented at the same time. If an infant consistently tended to look at one stimulus for longer than the others, the selectivity was thought to demonstrate the existence of perceptual discrimination.[1]

Fantz showed infants between the ages of four days and five months head-shaped discs resembling those in Figure 1 below. Infants of all ages spent significantly more time looking at the realistic face than at either of the other two. On the basis of this and other similar studies, Fantz determined that infants see a patterned and organized world, which they explore discriminatingly within the limited means at their command.

[1] Perceptual discrimination: the process of distinguishing differences in perceptions based on physical sensations.

Figure 1

Eleanor Gibson and Richard Walk also argued that infants possess well-developed perceptual skills. For their research, they designed a "visual cliff" (see Figure 2), which was actually a glass-top table. A check pattern was positioned close to the glass under one half of the table (the shallow side) and far below the glass under the other half (the deep side). Infants between the ages of 6 1/2 and 12 months were placed on the shallow side of the table and encouraged to crawl over the edge of the visual cliff onto the deep side by being offered toys or having their mothers call them. A majority of the infants failed to respond to these incentives, presumably because they possessed at least some of the elements of depth perception.

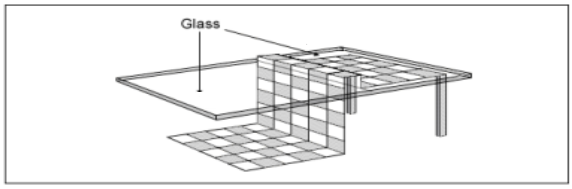


Figure 2

This work on the visual cliff, however, does not necessarily indicate that depth perception is innate, because infants who are several months old might have learned about depth perception from experience. The study did produce some intriguing physiological evidence pointing to the importance of learning in the visual cliff situation. Nine-month-old infants had faster heart rates than normal when placed on the deep side, suggesting that they were frightened. Younger infants of two months, on the other hand, actually had slower heart rates than usual when placed on the deep side, suggesting that they did not perceive depth and so were unafraid.

The question of the relative importance of innate factors and learning in perceptual development has not yet been resolved. It does appear probable, however, that innate factors and learning are both essential to normal perceptual development. Some of the basic elements of perception (e.g., perception of movement) seem to be either innate or else acquired very quickly.

In contrast, fine perceptual discriminations among objects (e.g., the ability to distinguish visually between similar letters such as “b” and “d”) may require much learning.

We can conclude that while innate factors provide some of the building blocks of perception, the complex perceptual processing of which adults are capable is learned over years of experience with the perceptual world.

Innate Versus Learned Perception

Speaking Question

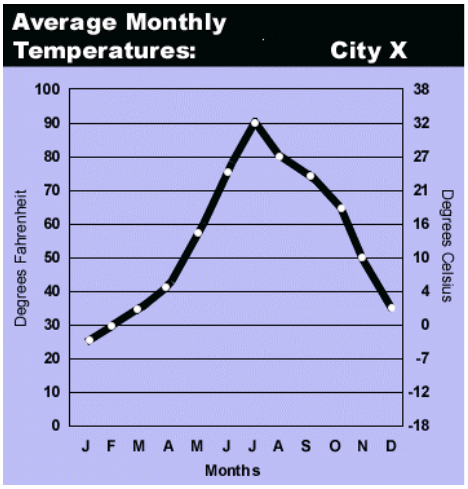
Describe Robert Fantz’s experiment in which he used the visual stimuli below.

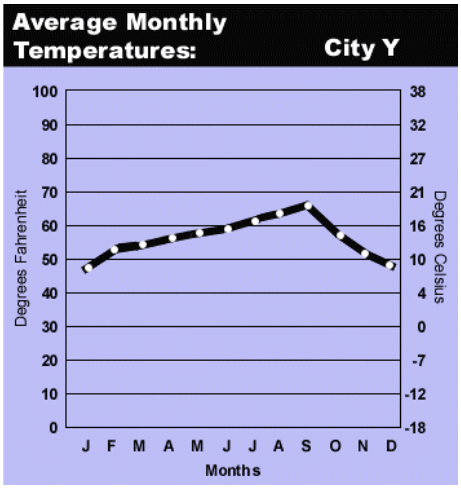
In your response, you should include information about

- the purpose of the experiment
- important details of the experiment

Sample Independent Speaking Tasks

1. A lot of events take place outside of the classroom at high schools and universities, such as athletic contests, student government, and music and drama performances. Please tell me in detail about an extracurricular event that happened within the past year at your high school or university that you were particularly interested or involved in.
2. The graphs below represent monthly temperatures in two cities in the United States. Based on the information given in the graphs, discuss the differences the climate might make on the lifestyles of people living in these two cities.


Average Monthly Temperatures: City X
Degrees Fahrenheit
Degrees Celsius
100 90 80 70 60 50 40 30 20 10 0
38 32 27 21 16 10 4 0 -7 -12 -18
J F M A M J J A S O N D
Months


Average Monthly Temperatures: City Y
Degrees Fahrenheit
Degrees Celsius
100 90 80 70 60 50 40 30 20 10 0
38 32 27 21 16 10 4 0 -7 -12 -18
J F M A M J J A S O N D
Months

Appendix D

Scoring Rubrics for Integrated and Independent Speaking Tasks

Integrated Speaking Task Scoring Rubric

5 A response at this level

- reports most major ideas *and* important supporting details accurately, though may have minor errors or omissions
- is marked by original, generally fluid speech and organized, complete thoughts; (may incorporate occasional verbatim phrases from source text)
- is marked by highly intelligible speech and requires little listener effort
- demonstrates appropriate use of a wide range of vocabulary and grammar; may have occasional grammatical or lexical errors

4 A response at this level

- reports many major ideas *and* important supporting details accurately, though may have omissions or exhibit some error in detail
- contains original, often continuous speech and complete thoughts, though may have some hesitation or lack of fluency (may occasionally use verbatim phrases from source text)
- is generally intelligible, though may require some listener effort
- demonstrates generally appropriate use of a good range of vocabulary and grammar, though may have minor errors and some imprecision and/or unidiomatic language use

3 A response at this level

- states/identifies some major ideas *and* important supporting details, though some information may be incomplete, inaccurate, or unclear
- contains some original and continuous speech and some complete thoughts, but may have considerable hesitation and/or disjunction(may use verbatim portions of source text)
- is reasonably intelligible with some listener effort needed throughout
- demonstrates at least some command of vocabulary and grammar, though problems with vocabulary and grammar occur regularly and may obscure meaning

2 A response at this level

- identifies a few major ideas *or* supporting details, but has errors and/or omissions of major points and details
- contains little original and continuous speech and few complete thoughts, with frequent hesitation and/or disjunction(may rely heavily on verbatim portions of source text)
- is somewhat intelligible, but only with considerable listener effort and has pronunciation errors that may interfere with intelligibility
- has very limited command of vocabulary and grammar; may have consistent errors in word choice and language that make it difficult to comprehend intended meaning

1 A response at this level

- mentions few or no important ideas required by task or does not address the task
- contains almost no continuous or original speech or complete thoughts, but with frequent hesitation, long pauses, and severe disjunction (or relies largely on verbatim portions of source text)
- may be substantially unintelligible and requires constant listener effort and shared knowledge for listener to comprehend

Independent Speaking Task Scoring Rubric

5 A response at this level

- develops a clear viewpoint about the topic, using well-chosen reasons and/or examples
- is marked by generally fluid speech and organized, complete thoughts
- is marked by highly intelligible speech and requires little listener effort
- demonstrates *control of* a wide range of vocabulary and grammar *with only minor or* occasional errors

4 A response at this level

- expresses clear thoughts or ideas (or opinion) about the topic, using some supporting reasons and/or examples
- contains original, often continuous speech and complete thoughts, though may have some hesitation or lack of fluency
- is generally intelligible, though may require some listener effort
- demonstrates generally appropriate use of a good range of vocabulary and grammar, *with* minor errors and some imprecision and/or unidiomatic language use

3 A response at this level

- addresses the topic, but development may be limited or unclear
- contains at least some continuous speech and some complete thoughts, but may also have considerable hesitation and/or disjunction;
- is reasonably intelligible, but requires some listener effort throughout
- demonstrates at least some command of vocabulary and grammar, though problems with vocabulary and grammar occur regularly and may obscure meaning

2 A response at this level

- attempts to address the task but offers little clear or understandable development
- contains little continuous speech and few, if any, complete thoughts, and has frequent hesitation and/or disjunction (may rely heavily on repetition of the prompt)
- is intermittently intelligible, but only with considerable listener effort and has pronunciation errors that interfere with intelligibility
- has very limited command of vocabulary and grammar; consistently has errors in word choice and language that make it difficult to comprehend intended meaning

1 A response at this level

- does not provide appropriate response to the topic or provides only very basic thoughts about the topic beyond a restatement of the prompt
- contains almost no continuous speech and has frequent hesitation, long pauses, and severe disjunction (or relies largely on verbatim portions of prompt)
- may be mostly unintelligible and requires constant listener effort and shared knowledge for listener to comprehend

Test of English as a Foreign Language
PO Box 6155
Princeton, NJ 08541-6155
USA

To obtain more information about TOEFL programs and services, use one of the following:

Phone: 1-877-863-3546
(US, US Territories*, and Canada)

1-609-771-7100
(all other locations)

Email: toefl@ets.org

Web site: www.ets.org/toefl

* America Samoa, Guam, Puerto Rico, and US Virgin Islands

I.N. 726812