ETS

TOEFL

ISSN 1930-9317

TOEFL iBT Research Report

TOEFLiBT-01
April 2006

Investigating the Utility of Analytic Scoring for the TOEFL Academic Speaking Test (TAST)

Xiaoming Xi

Pam Mollaun

Listening.
Learning.
Leading.

Investigating the Utility of Analytic Scoring for the TOEFL Academic Speaking Test (TAST)

Xiaoming Xi and Pam Mollaun

ETS, Princeton, NJ

RR-06-07

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.

Abstract

This study explores the utility of analytic scoring for the TOEFL® Academic Speaking Test (TAST) in providing useful and reliable diagnostic information in three aspects of candidates' performance: delivery, language use, and topic development.

G studies were used to investigate the dependability of the analytic scores, the distinctness of the analytic dimensions, and the variability of analytic score profiles. Raters' perceptions of dimension separability were also obtained.

Based on the phi coefficients and standard errors of measurement (SEMs), the dependability of analytic scores averaged across six tasks and double ratings was acceptable for both operational and practice settings. However, scores averaged across two tasks and double ratings were not reliable enough for operational use.

Correlations among the analytic scores by task were high, but those between delivery and topic development were lower, and these results were corroborated by raters' perceptions. When averaged across tasks or task types (two or more tasks), correlations among the analytic scores were very high, and the profiles of scores were flat.

The utility of analytic scoring is discussed, and both score dependability and whether analytic scores provide diagnostic information beyond that provided by holistic scores are considered.

Key words: TOEFL iBT speaking, analytic scoring, score dependability, dimension distinctness, score profile, G theory

The Test of English as a Foreign Language™ (TOEFL®) was developed in 1963 by the National Council on the Testing of English as a Foreign Language. The Council was formed through the cooperative effort of more than 30 public and private organizations concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS®) and the College Board® assumed joint responsibility for the program. In 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations® (GRE®) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a policy board that was established by, and is affiliated with, the sponsoring organizations. Members of the TOEFL Board (previously the Policy Council) represent the College Board, the GRE Board, and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.

❖ ❖ ❖

Since its inception in 1963, the TOEFL has evolved from a paper-based test to a computer-based test and, in 2005, to an Internet-based test, TOEFL iBT. One constant throughout this evolution has been a continuing program of research related to the TOEFL test. From 1977 to 2005, nearly 100 research and technical reports on the early versions of TOEFL were published. In 1997, a monograph series that laid the groundwork for the development of TOEFL iBT was launched. With the release of TOEFL iBT, a TOEFL iBT report series has been introduced.

Currently this research is carried out in consultation with the TOEFL Committee of Examiners. Its members include representatives of the TOEFL Board and distinguished English as a second language specialists from the academic community. The Committee advises the TOEFL program about research needs and, through the research subcommittee, solicits, reviews, and approves proposals for funding and reports for publication. Members of the Committee of Examiners serve four-year terms at the invitation of the Board; the chair of the committee serves on the Board.

Current (2005-2006) members of the TOEFL Committee of Examiners are:

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

Acknowledgments

We would like to thank all of the assessment development specialists who participated in the analytic scoring study, with special thanks to Rob French for contributing his expertise and experience to this project and for actually conducting some of the analytic scoring training sessions. We are grateful to Emily Midouhas for her various research assistance on this project, particularly for summarizing the rater survey data and for carefully proofreading an earlier version of this manuscript. Our thanks also go to Kitty Sheehan, who obtained the references needed for this project and helped prepare the manuscript.

Table of Contents

Page

List of Tables

Page

Context and Purpose

TOEFL® Academic Speaking Test (TAST) is a speaking assessment developed to measure examinees' ability to communicate orally in an academic environment. Its intended use is as an indicator of the adequacy of candidates' oral communication skills for studying in colleges and universities in English-speaking countries. In September 2005, when the TOEFL Internet-based Test (TOEFL iBT) launched in the United States, TAST was included as the speaking section. TAST is currently available as a stand-alone test used by individuals to practice for TOEFL iBT. Since TAST has been designed to engage the essential speaking skills required in academic settings (Butler, Eignor, Jones, McNamara, & Suomi, 2000; Douglas, 1997; Rosenfeld, Leung, & Oltman, 2001; Waters, 1996), it can potentially be used for placing students into different types and levels of remedial English classes.

The test consists of six speaking tasks. Two of them integrate listening and speaking skills, and two integrate reading, listening, and speaking. The content of the reading and listening materials reflects that found in student life experiences and basic academic coursework. In the remaining two tasks, speakers are asked to provide information or opinions on familiar topics based on their personal experience or background knowledge.

TAST is currently rated using a holistic rubric. With the TAST holistic scale, raters evaluate examinees' overall performance on each task by considering the combined impact of their delivery, language use, and topic development (see Appendix A for the TAST scoring rubric). An alternative scoring approach, analytic scoring, can be used to assess examinees' performance on each of the three dimensions. In other words, separate delivery, language use, and topic development scores can be reported.

In language testing, the debate between holistic (or global) and analytic (or componential) rating rubrics for speaking tests has been well-documented (Bachman, 1988; Bachman & Savignon, 1986; Douglas & Smith, 1997; Fulcher, 1997; Ingram & Wylie, 1993; Underhill, 1987; Weir, 1990). Holistic scoring is often preferred over analytic scoring for oral tests that attempt to evaluate the overall communicative effectiveness of the candidates (Weir. In addition, holistic scoring promises efficiency in scoring and ease in score reporting and is likely to impose a lesser cognitive load on raters. However, it also poses some potential problems.

First, in holistic scoring the relative weights of the subfeatures defined in the scoring rubric are implicit and may also be idiosyncratic. The language components or dimensions in the holistic rubrics may be weighted differentially by different raters depending on their background and experience and their perceptions of how a particular weakness or strength impacts the overall communicative quality in a particular assessment context (Brown, 1995). Even though there are similarities in how different raters derive a holistic score for a specific response, there are often few explicit rules that raters can utilize when making a global judgment. Analytic scoring makes a systematic way of weighting different dimensions based on the assessment focus for each task type possible.

Another problem with holistic scoring is related to the interpretation of holistic scores. As Weir (1990) points out, the typical performance descriptions at each holistic score level might not work for candidates who demonstrate varied performance on the various components. In the context of the TAST, although a single holistic score might be sufficient for making admission decisions, decisions regarding placing students into different types and levels of remedial classes will require more diagnostic information (i.e., what a particular examinee's strengths and weaknesses are). The institutions will want to use the diagnostic information to place students into different remedial speaking classes and to adjust their class activities to match students' strengths and weaknesses. Examinees may wish to use the diagnostic information to guide their language learning activities. With the holistic rating rubric, it is possible to generate a score report that describes the performance of examinees at particular score levels. However, the descriptors at each score level in the holistic rubric intend to capture *typical* profiles of performance (performing equally well on the different dimensions) and do not take into account the variety of profiles that might exist within the holistic score levels. In other words, examinees who receive the same holistic score may have very different profiles due to unequal performance on the dimensions considered globally in holistic scoring. The descriptions of typical performance at each holistic score level may not capture the performance profiles of these examinees. Score reports based on holistic scoring may provide only limited information about the candidate's performance, while those based on analytic scoring are able to capture more varied profiles.

Analytic scores can provide this kind of diagnostic information for examinees with varied profiles (Bachman & Savignon, 1986), while allowing for the possibility of generating a single composite holistic score if proper weights are applied to the dimensions. A score report generated from the analytic ratings thus has the potential to serve both admissions purpose (where a single score is desired) and placement purpose (where more diagnostic information is needed). In addition, Bachman and Savignon also argue that speaking ability is a multicomponential trait and that rating rubrics should be defined in terms of components such as functional, grammatical, discourse, and sociolinguistic competencies, to reflect current models of communicative language ability.

Despite its appeal as a preferable scoring approach, analytic scoring is not without its problems. Potential rating inconsistencies due to high cognitive load on raters, difficulty in defining the dimensions in analytic rubrics precisely, and difficulty calibrating raters have been most frequently mentioned (Douglas & Smith, 1997; Underhill, 1987). Underhill notes the difficulty raters experienced when they had to evaluate the candidate's performance on several criteria at the same time. Douglas and Smith suggest that the use of a holistic rating system on the Test of Spoken English™ (TSE®) allows for more consistent ratings. They argue that it is very difficult to define the components so that each rater agrees on the precise meanings and feels comfortable assigning component ratings. Raters may feel more comfortable working with a certain degree of "fuzziness" when using a holistic rubric.

In order for analytic scoring to be useful for an assessment, raters must be trained to reliably distinguish among the dimensions. In addition, the target population must demonstrate sufficiently varied profiles to warrant a more costly and complex analytic scoring system.

This study attempts to explore empirically the utility of analytic scoring from the perspectives of score dependability, dimension separability, and rater perceptions. The purpose of this study is to investigate the utility of analytic scoring for TAST (or TOEFL iBT speaking) in large-scale operational settings as well as learning and practice environments, with the focus on its utility for operational use.

Specifically, this study intends to answer the following questions:

1. To what extent are the analytic scores dependable?
2. How is the dependability of composite scores impacted if different weighting schemes are used?
3. To what extent are the dimensions separable?
4. To what extent are examinees' profiles varied?

5. What is the relationship between the holistic and analytic scores? Is the relationship similar for independent and integrated tasks?

The G Theory Framework

Univariate G Theory

In this study, generalizability (G) analyses were employed to investigate the dependability of the analytic scores and the composite scores, which are averages or weighted averages of the analytic scores. In the classical test theory (CTT) framework, the observed score is defined as the universe score plus error, where error is a single value and cannot be decomposed. In CTT, interrater reliability and Cronbach Alpha are used to estimate rater reliability and task reliability respectively. When interrater reliability is estimated, it is assumed that scores are averaged over an infinite number of tasks and that there is no sampling variability due to tasks. Similarly, when Cronbach Alpha is used to estimate the internal consistency of test tasks, it is assumed that raters are perfectly consistent and that no variability is associated with them. In both cases, the variance associated with tasks or raters actually goes into the universe score variance, thus the universe score variance is overestimated. In the G theory framework, different sources of error can be estimated simultaneously, rather than separately as in CTT. The magnitude of each specified source of error can be estimated, along with the amount of universe score variance.

A G analysis is conducted in two stages. In the first stage, or G study, variance components based on the sample data are estimated for different sources of score variation (the object of measurement, facets and interactions of facets). The sources of variation include main effects and interaction effects, following the terminology for standard analysis of variance. The object of measurement, usually persons (*p*), is associated with a main effect and indicates the extent to which observed variance in persons' scores is due to real differences in ability levels. Main effects associated with facets demonstrate the extent to which averaged scores are the same across different levels of the facets, whereas interaction effects indicate consistency in the rank ordering of examinees. For example, in a G study design where persons (*p*) is the object of measurement and raters and tasks are the two facets, the rater main effect indicates raters' leniency or harshness in their judgments (i.e., to what extent the mean scores assigned by different raters across tasks and persons are the same). On the other hand, the person-by-rater interaction is an indication of the extent to which persons (*p*) are rank ordered similarly by different raters. Variance components are estimated for each main effect and interaction effect specified in the G study. These variance components obtained from a G study are assumed to be generalizable to a universe of defined situations. Variance components obtained from the G study serve two major purposes. First, they provide the basis for the subsequent Decision (D) studies that aim to find an optimal assessment design; second, they pinpoint the aspects of assessment that need to be improved.

In the second stage, or D studies, the variance components from the G study are used to estimate variance components and generalizability coefficients (G and phi coefficients) for alternative measurement designs where the levels of the facets are varied. It is assumed that all the levels of the facets, which are randomly drawn from an infinite universe of generalization, are randomly parallel. It should be noted that variance components obtained in a G study are based on scores for a single observation, for example, when a single task is used and a single rating is obtained for each person on each task. In practice, examinees' scores are typically based on multiple tasks and multiple ratings. Increasing the number of raters and tasks would reduce the magnitude of all variance components except the one for persons (*p*), the true variance. For example, if two tasks are used and double ratings are obtained for each examinee on each task, the variance components for the rater main effect and the task main effect are expected to each reduce by half. The other variance components are expected to reduce correspondingly as well. Because the raters and tasks are assumed to be randomly parallel (i.e., exchangeable with other raters and tasks in the universe), the basic idea is that the more tasks and ratings per task that examinees' scores are based on, the smaller the variance components for errors and the closer a typical examinee's observed score will be to his/her universe score.

The D study variance components are used to estimate error variances and G and phi coefficients, which show the proportion of universe score variance to observed score variance for two different kinds of decisions. The observed score variance equals the sum of the universe score variance and error variance. The magnitude of error is dependent on the type of decisions based on scores. When decisions are based on relative standings of examinees, such as selecting the top performing examinees, the error variance consists of interaction effects only. This is called relative error variance. However, when decisions are based on the absolute values of scores, such as determining examinees' levels of performance as compared to a criterion, the error variance, called absolute error variance, is the sum of all variance components (both main and interaction effects) except the one for persons (*p*).

By the same token, the G coefficient indicates how reliable the scores are when one is only concerned with how examinees are rank ordered compared to others, for example, how consistent examinees' relative standings are across raters, tasks, or rater-task combinations. The phi coefficient applies to situations where the concern is for the absolute value of scores, for example, how consistently an examinee will earn a specific score across different raters, tasks, or rater-task combinations.

Similarly, there are two types of SEM, one for relative decisions and one for absolute decisions. It is the square root of total error variance, be it relative or absolute error variance. SEMs for relative decisions in G theory are analogous to SEMs in CTT, but SEMs for absolute decisions cannot be estimated in the framework of CTT.

SEM is more informative than G and phi coefficients for decision-making (Brennan, Gao, & Colton, 1995; Linn & Burton, 1994) and is also easier to understand and interpret conceptually since it is expressed on the same scale as the scores. Both G and phi coefficients depend on universe score variance and indicate the magnitude of true differences among persons (universe score variance) relative to errors. Thus, if universe score variance is small relative to error variance, G and phi coefficients could be small even if error variances are small. Therefore, it may not be sensible to determine the number of tasks and ratings based on the magnitude of these coefficients alone (Brennan, 2002). SEM, on the other hand, shows to what extent a typical examinee is accurately measured given a specific assessment context, but it should be treated as a rough approximation since its normality assumption is not always satisfied in practice (Brennan 2000; Brennan et al., 1995). G and phi coefficients provide overlapping but different types of information about the precision of examinees' scores.

Since TOEFL iBT speaking is intended to be used for admission and placement purposes (where applicants' scores are compared to predetermined cut scores), which may vary by institution and program, the users will need to know how dependable candidates' absolute scores are (the phi coefficient). These users would be primarily interested in absolute decisions, that is, using the absolute scores to gauge a candidate's readiness for studying at the post-secondary level in English-speaking countries or to determine a candidate's placement level in speaking class. They would be particularly interested in the dependability of examinees' scores around the cut scores, since these examinees are the people whose chance of getting admitted or getting placed into a class would most likely be affected by errors in scoring. It is possible to use phi lambda to estimate the dependability of scores given a particular cut score (Brennan, 2001), and it would be necessary to report this index if the cut scores were known. The magnitude of phi lambda depends on the cut score. When the cut score is equal to the mean of all scores, the estimate of phi lambda reaches its minimum value (Brennan).

Moreover, the different forms of TAST are not equated. Although rigorous processes are followed in task development, scoring rubric development, and rater training to ensure the comparability of forms, minor differences in difficulty across forms may still exist. In this case, phi coefficients are more appropriate than G coefficients since institutions are comparing applicant scores that may be based on different forms.

Multivariate G Theory

Unlike univariate G theory, in which each object of measurement (usually *persons*) has one universe score, in multivariate G theory each object of measurement has multiple universe scores. Thus a multivariate G analysis decomposes both variances and covariances among universe scores and among errors into components (Brennan, 1992, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1981; Webb, Shavelson, & Maddahian, 1983). It is appropriate for analyzing the dependability of composite scores on tests that measure multiple content domains or traits that are related.

The covariance components provide some additional information about how examinees' universe scores and errors co-vary. The covariance components for different effects can be interpreted relative to the corresponding variance components for the multiple universe scores.

An important aspect of the development of multivariate G theory is the distinction between linked and unlinked conditions. The expected values of error covariance components are nonzero when conditions are linked or jointly sampled (e.g., ratings on all three dimensions, delivery, language use, and topic development are assigned by the same raters). These conditions are indicated by a filled circle (e.g., $r^{\bullet}$). The expected values of error covariance components are zero when conditions for observing different variables are unlinked or selected independently (e.g., the raters on one dimension, delivery, are different than those on language use or topic development). Unlinked conditions are represented by an unfilled circle (e.g., r°).

Multivariate G theory has two important applications: one is to estimate the dependability of composite scores using different weighting schemes. One can explore how weighting the multiple measures differentially can impact the dependability of the composite score (Brennan, 2001; Joe & Woodward, 1976; Marcoulides, 1994). The other frequently reported result of a multivariate G study is the correlations among the universe scores that reveal the "true" relationships among scores on multiple measures corrected for score unreliability.

Method

Sample

One hundred eighty nonnative speakers studying in seven U.S. universities participated in a field study of the TAST. The sample was recruited in a way that ensured that various English proficiency levels, native language backgrounds, and educational experiences were represented (see Appendix B).The sample included both students who were currently enrolled in colleges or graduate schools and those in intensive English programs preparing to attend colleges.

Of these 180 examinees, 100 had tasks that were completely double scored and 80 had tasks that were single scored using the holistic scoring rubric. All nonadjacent discrepancies were adjudicated. These 100 examinees in the double scored group represented a range of proficiency levels. Each task was rated on a scale of 0-4 with 0 indicating no response or no attempt to respond. Holistic scores on individual tasks were averaged to yield the final overall holistic scores.

Following this, a stratified sample of 140 examinees was selected from this sample of 180 to be analytically scored. The number of examinees scored at each holistic score range was proportional to that in the 180-person sample and frequency counts were also obtained on the native language backgrounds of this 140-person sample to make sure that major native language groups were well-represented. A total of 34 native languages were represented in this sample (see Appendix C). Over a third, 34% of them, were enrolled in intensive English classes only, and the rest were ESL students enrolled in content classes in those seven U.S. universities.

Seventy-nine of these 140 examinees had been double rated holistically (see Appendix D for the descriptives of the holistic scores).

Material

A single TAST test form was used in this study. The test contains six speaking tasks. The first two tasks ask the examinees to speak about familiar topics; these are independent tasks. The remaining four tasks are integrated tasks, and the examinees must use more than one skill when responding to these tasks. Tasks 3 and 4 integrate speaking with listening and reading. One task involves a campus-based situation, and the other involves an academic topic. Tasks 5 and 6 integrate listening and speaking, and one task is campus-based while the other is academic. The listening and reading materials are short. To take the test, examinees printed a paper booklet from a designated web site and dialed in via an interactive voice response (IVR) system. They could take notes to use when responding to the speaking tasks. The test was approximately 20 minutes long. For each of the six questions, examinees were given a short time to prepare a response. The response time allowed for each question ranged from 45 to 60 seconds.

Holistic and Analytic Scoring Rubrics

Prior to this study, a scoring study was conducted to develop, evaluate, and refine holistic scoring rubrics for the TAST. In the first phase of the study, using the data gathered for the study, described earlier, the responses to each of the six tasks were analyzed by a group of individuals with varied backgrounds in applied linguistics and/or teaching English as a second/foreign language who rank ordered the responses and identified salient features. Then the assessment specialists who developed the TAST tasks collated these features for different levels of performance. Four band levels and three key categories of performance features emerged: delivery, language use, and topic development. Descriptions of features in each category and at each level were formulated and draft holistic rating scales were devised from the formulations. In the second phase of the study, the responses were scored using the draft holistic scales. The results were analyzed to investigate use of the holistic scales by raters and score distributions across band levels. The next step was to investigate scoring strategies other than *holistic*, using the key dimensions identified in the scoring study.

The descriptors for the four levels (1-4) of delivery (D), language use, (L) and topic development (T) from the TAST holistic scoring rubric were used to create a separate analytic rubric for each dimension (Appendix A). The same 1-4 scale was adopted for each of the analytic rubrics. As specified in the holistic rubric, raters were instructed to follow these guidelines during holistic scoring: an examinee must be on target for all the three dimensions to receive a score of 4 and for at least two of the dimensions to receive a score of 1, 2, or 3. These guidelines helped raters make overall holistic judgments.

Delivery refers to the pace and clarity of the speech. In assessing delivery, raters consider the speakers' pronunciation, intonation, rate of speech, and degree of hesitancy. Language use refers to the range, complexity, precision, and automaticity of vocabulary and grammar use. Raters evaluate candidates' ability to select words and phrases and to produce structures that appropriately and effectively communicate their ideas. Topic development refers to the coherence and fullness of the response. When assessing this dimension, raters take into account the progression of ideas, the degree of elaboration, the completeness, and, in integrated tasks, the accuracy of the content.

Design of the Analytic Rating Sessions

All six responses from these 140 examinees were rated on each of these three dimensions. Table 1 shows the means and standard deviations of the three analytic scores by task of the 140-person sample. Overall, the means of different dimensions and of different task-dimension combinations were quite close.

The analytic scoring was conducted in two phases. In the first phase (Table 2), 30 examinees' responses to Tasks 2, 4, and 5 were quadruple rated by four raters (A-D) on all three dimensions. These 30 examinees represented a stratified sample with varied proficiency levels and native language backgrounds. The training for raters on the three dimensions was conducted separately; they rated one dimension and one task at a time. The examinees were scrambled each time a new dimension or a new task was rated, and they were also scrambled when rated by a different rater. The purpose of conducting this Phase 1 rating was to test the usability of the analytic rating scales and to examine whether obtaining double ratings for each task would yield reasonably high score dependability. Once the results from the Phase 1 rating confirmed that the dependability of analytic scores with double ratings was acceptable, Phase 2 rating was conducted.

In Phase 2 (Table 3), 14 raters (A-N) double rated 140 examinees on Tasks 1, 3, and 6 and 110 examinees[1] on Tasks 2, 4, and 5 on all dimensions, one task at a time and one dimension at a time, as done in the first phase. The examinees were randomly divided into four blocks of 30, 37, 37, and 36. As shown in Table 3, a different rater pair rated each dimension of task sets 1, 3, and 6 or 2, 4, and 5 for each examinee block.

Table 1

Descriptives of the Analytic Scores of the 140-Person Sample

	Dimension	Mean	Std. deviation
Task 1	D	2.52	1.02
	L	2.25	0.95
	T	2.39	1.07
Task 2	D	2.56	1.02
	L	2.42	0.95
	T	2.53	1.01
Task 3	D	2.49	0.87
	L	2.39	0.93
	T	2.32	1.01
Task 4	D	2.49	1.02
	L	2.38	0.95
	T	2.37	1.02
Task 5	D	2.61	1.03
	L	2.42	0.98
	T	2.44	1.05
Task 6	D	2.53	0.88
	L	2.31	0.86
	T	2.53	0.94
All tasks	D	2.53	0.86
	L	2.36	0.83
	T	2.43	0.86

Note. D = delivery, L = language use, T = topic development.

Table 2

Phase 1 Rating Design

Examinee ID	Tasks	Delivery	Language use	Topic development
1-30 (30)	2, 4, and 5		*A, B, C, D*	

Table 3

Phase 2 Rating Design

Examinee ID	Tasks	Delivery		Language use		Topic development	
		Rater 1	Rater 2	Rater 1	Rater 2	Rater 1	Rater 2
1-30 (30)	1, 3, 6	E	H	G	M	F	K
31-67 (37)		A	I	D	N	G	L
68-104 (37)		B	J	A	H	E	M
105-140 (36)		C	K	B	I	A	N
31-67 (37)	2, 4, 5	D	L	C	J	B	H
68-104 (37)		F	M	E	K	C	I
105-140 (36)		G	N	F	L	D	J

Rater Questionnaire

After the analytic scoring was completed, all the raters (N = 14) filled out a questionnaire; they provided their background information and reflected on their analytic scoring experience (E). They reported their teaching experience in English as a second language or English as a foreign language (ESL/EFL), degrees they have obtained, and the foreign languages they are familiar with, among other things. They also rated on a scale of 1-4 the ease/difficulty with which they could understand heavily accented speakers whose native languages they were familiar with (from "very easily" to "very difficult").

The focus of the questionnaire was on the extent to which the raters were able to distinguish among the three dimensions. They rated on a 1-4 point scale how much confidence they had in rating each dimension ("not at all confident" to "very confident") and how much overlap they thought there was among the three dimensions ("very distinct" to "almost impossible to distinguish"). They also rated the extent to which they agreed or disagreed with certain statements designed to measure how well raters thought they were able to tease apart the three dimensions.

For example, one of such statements is: *"When rating a speaker with a strong accent, I listened a few times to rate his/her language use."* The raters indicated their reactions to this statement on a 1-6 scale from "strongly disagree" to "strongly agree."

Rater Characteristics and Qualifications

Seventy-nine percent of the raters are native speakers of English, while the remaining raters are fluent bilingual speakers of English and another language. All are involved in test development for English language learning assessments (ELL) programs, and 57% indicated that they have had experience teaching ESL/EFL classes either in the United States or overseas. Those with ESL/EFL teaching experience taught for an average of 12 years, and 79% of the raters have specialties in a language field (e.g., applied linguistics, TESOL, modern languages).

Analyses and Results

Rater Agreement

Rater agreement rate. Table 4 shows the rater agreement rate by dimension and by task. There were no noticeable differences in rater agreement rates across delivery, language use, and topic development; the combined exact and adjacent agreement rates were similar for the three dimensions. The exact agreement rates were not problematic; however, the nonadjacent agreement rates for language use and topic development were a little high (4.7% and 5.5%), compared to other large-scale speaking assessments such as the TSE, which uses holistic scoring. Some of the largest percentages of nonadjacent discrepancies occurred with topic development scores on Tasks 2 and 6. The nonadjacent discrepancies in topic development scores for Task 2 were mainly associated with one rater pair, which accounted for 10 of the 13 discrepancies. Using the adjudicated scores as the criteria, one of the two raters rated consistently more leniently than is appropriate.

The nonadjacent discrepancies associated with topic development scores on Task 6 were spread out across the four rater pairs assigned for this task. This listening/speaking task was identified by the trainers of the raters as somewhat problematic in that the relationships among the major concepts in the audio stimulus material were not particularly well-marked. Examinees therefore interpreted the stimulus inconsistently. While guidelines were provided for raters to judge the responses, some raters may have had difficulty determining the appropriateness of the response to the task.

An unexpected result was a 10% nonadjacent discrepancy rate associated with delivery ratings on Task 2. A close examination revealed, again, that more than half of the discrepancies were produced by one rater pair. In particular, one of the raters was consistently more lenient on delivery than the other, especially with speakers of a particular native language background. This particular rater indicated on the questionnaire that it was relatively easy for him/her to understand heavily-accented speakers of this language (2 on a scale of 1-4 with 1 indicating "very easy" and 4 suggesting "very difficult"). It also turned out that this particular rater did not have any ESL/EFL teaching experience.

All the nonadjacent discrepancies were adjudicated by the lead person in rubric development and rater training. The adjudicated scores were used in the subsequent G study analyses since examinees' final scores were based on adjudicated scores.

Table 4

Agreement Rate Between First and Second Ratings on the Three Dimensions by Task

Task	# of ratings	Delivery			Language use			Topic development		
		Exact	Adj.	Non.	Exact	Adj.	Non.	Exact	Adj.	Non.
1	140	55.7%	44.3%	0.0%	49.3%	42.1%	8.6%	52.1%	42.9%	5.0%
2	110	43.6%	46.4%	10.0%	52.7%	44.5%	2.7%	43.6%	44.5%	11.8%
3	140	69.3%	29.3%	1.4%	49.3%	42.9%	7.9%	50.7%	45.7%	3.6%
4	110	51.8%	46.4%	1.8%	60.9%	36.4%	2.7%	72.7%	26.4%	0.9%
5	110	55.5%	40.0%	4.5%	67.3%	30.9%	1.8%	51.8%	43.6%	4.5%
6	140	69.3%	29.3%	1.4%	49.3%	47.9%	2.9%	46.4%	46.4%	7.1%
Total	750	58.4%	38.7%	2.9%	54.1%	41.2%	4.7%	52.5%	42.0%	5.5%

Note. Adj. = adjacent, Non. = nonadjacent.

Raters' self-reported confidence in rating the three dimensions. In the rater survey, all of the raters reported on a scale of 1-4 the level of confidence they had in rating each of the three dimensions with 1 indicating "Not at all confident" and 4 "Very confident." As illustrated in Table 5, 12 out of the 14 raters indicated that they were confident or very confident in rating language use and topic development and 10 indicated confidence in rating delivery (a rating of "3" or "4"). None of the raters without ESL/EFL teaching experiences felt "Very confident" in rating the three dimensions, and they lacked confidence in rating delivery particularly.

The reasons raters cited for their confidence or lack of confidence in rating a particular dimension had much to do with the degree to which they felt they could tease apart a particular dimension, while blocking out interference from other dimensions.

Table 5

Raters' Confidence in Rating Delivery, Language Use, and Topic Development

Level of confidence in rating	Delivery		Language use		Topic development	
	Freq.	Perc.	Freq.	Perc.	Freq.	Perc.
Not at all confident (1)	0	0%	0	0%	0	0%
Somewhat confident (2)	4	29%	2	14%	2	14%
Confident (3)	6	43%	8	57%	9	64%
Very confident (4)	4	29%	4	29%	3	21%

The reasons given for not being confident in rating delivery included lack of ESL/EFL experience, difficulty in classifying an infinite variety of responses into four distinct delivery categories, difficulty in distinguishing a 2 and a 3 rating, inability to separate delivery from poor grammar and content, having a nonnative ear, an so on All but one rater with ESL/EFL teaching background reported being "Confident" or "Very Confident" in rating delivery. The reasons they cited were ESL/EFL experience, well-run training sessions, ease in using the delivery rubric, fruitful discussions, and ease in separating delivery from other dimensions, among other things. The only rater with ESL/EFL experience who indicated a lack of confidence in rating delivery felt it hard at times to determine whether someone's delivery indicated problems with the target language or a problem with the topic, about which the person may have had little to say.

Some of the raters who indicated less confidence in rating language use thought there was a significant amount of overlap to contend with between language use and the other two dimensions. Others indicated uncertainty about assessing examinees who have limited vocabulary and grammar resources but make effective use of them, which suggests that the raters had difficulty applying the rubrics. Quality calibration materials and training, experience, and ease in evaluating the range and complexity of grammar were among the reasons mentioned by raters who felt confident at rating language use.

Some raters did not feel confident rating topic development because they felt it was hard to distinguish between topic comprehension and topic development. Some thought topic development was impacted by delivery and language use, in that effective use of intonation markers and cohesive devices facilitates progression of ideas. Still others felt less confident and expressed ambiguity about whether the content criteria were clearly met. However, some raters reported confidence, and mentioned among the reasons good calibration materials, well-run training, and useful topic notes that outlined the major points that needed to be covered for each task.

Score Dependability Analysis

Univariate G Analyses on the Dependability of Analytic Scores

Phase 1 rating featured a fully crossed p x t x r design for each of the three dimensions of delivery, language use, and topic development. Phase 2 rating features an r: (p x t) design, with persons (*p*) crossed with tasks (*t*) and raters (*r*) nested within persons and tasks. However, this nested design, although commonly used in large-scale performance assessments, does not allow estimation of independent variance components involving raters such as *r, rt, pr,* or *ptr***.** Thus, very limited information could be obtained regarding errors associated with raters. So that all effects could be estimated, an alternative analysis was conducted to treat each rater pair as a block, estimate it as a fully crossed design, and then pool variances across the rater pairs (for references on pooling variance components, see Brennan et al., 1995; Chiu, 2000; Chiu & Wolfe, 2002; Smith, 1980; Wiley, 1992). This analysis allows us to use all the data in both Phase 1 and Phase 2 and to examine the impact of all sources of error on score dependability independently.

As Brennan et al. (1995) and Chiu (1999) have demonstrated, the variance and covariance components from multiple G studies can be pooled to obtain more accurate, stable, and comprehensive variance and covariance component estimates. These averaged variance and covariance components can then be used in D studies to yield more accurate estimates of score dependability given the alternative measurement designs specified. This method is especially preferable for scenarios where multiple rating schemes are utilized, as in the present study, in which different rating designs (different G study structures) were used in Phase 1 and Phase 2 (Chiu & Wolfe, 2002). Pooling variance and covariance components from all G studies would allow us to keep all the data in the analysis. Otherwise, tossing out valuable data could result in unstable and inaccurate variance and covariance component estimates (Chiu & Wolfe, 2002).

There were seven fully crossed p x t x r designs for delivery in Phase 2 rating, each row under Delivery in Table 3 representing one p x t x r design, where persons were crossed with tasks with raters. Phase 1 rating in Table 2 also included a fully crossed p x t x r design for delivery. Because each response was quadruple rated in Phase 1 and double rated in Phase 2, variance components associated with raters were estimated from a sample of four raters for the p x t x r design in Phase 1 and from a sample of two raters for each of the seven p x t x r designs in Phase 2. Nevertheless, they were all p x t x r designs.

Task sets 1, 3, and 6 and 2, 4, and 5 are parallel in that each set includes an independent task on a familiar topic, an integrated task on academic course content, and an integrated task on campus situations. Note that these eight designs were not strictly independent since some of them shared the same examinees.

Variance components were estimated for each of these eight designs for delivery and pooled together. Variance components for language use and topic development were pooled together in a similar fashion.[2] The estimated G study variance components for delivery, language use, and topic development are illustrated in F. In each table, estimated variance components for different rater pairs are reported separately, along with the average of the eight estimates and an estimate of the standard error (SE) for the average. The estimated SE is the standard deviation of the eight estimated variance components divided by the square root of eight.

The SEs of the averages, especially for these relatively large variance components (such as *p*, *pt,* and *ptr*), were generally small compared to the estimated variance components, indicating that the averages were fairly stable estimates of the variance components. In the subsequent D studies on the analytic scores, averaged variance components from these eight analyses were used. Table 6 shows the averaged variance components and the percentages of the total variance accounted for by each source of variation.

Examinees were almost equally variable in their delivery, language use, and topic development, as indicated by the similar variance components associated with persons (true variance in CTT) on these three dimensions. Overall, a substantial proportion of the total variance in the scores on the three dimensions could be explained by real differences in examinees' delivery, language use, or topic development. Among the three dimensions, 65.1% of the variance in examinees' delivery scores was explained by variance associated with persons, suggesting that examinees' delivery scores were the most reliable. For all dimensions except topic development, the largest source of error was the person-by-task-by-rater interaction and other undifferentiated errors. The person-by-task interaction was the next largest source of error for delivery and language use and the largest for topic development, suggesting that examinees were rank ordered very differently on each of these three dimensions across the tasks. That is to say, depending on which tasks examinees take, their relative standings in delivery, language use, or topic development may be different. The person-by-task interaction for topic development was the largest among the three dimensions, suggesting that the students' relative standings in topic development were the most varied across tasks. Given that topic development is the most task-specific feature in the rubrics, it was expected that examinees' topic development scores would be rank ordered very differently across the tasks. The rater main effects or the person-by-rater interactions were generally small, which means that raters did not differ much in their leniency or harshness or in judging where an examinee stood compared to other students.[3] The task main effects were almost zero, indicating that the mean scores of this group of examinees were the same across the tasks (i.e., on average, the tasks varied little in their difficulty levels). The t x r interaction was negligible, indicating minimal difference in raters' rank orderings of task difficulty**.**

Table 6

Variance Components for the p x t x r Design

Sources of variation	Variance component			Percent of total variation		
	D	L	T	D	L	T
P	.657	.604	.637	65.1%	61.5%	55.4%
T	.003	.006	.019	0.3%	0.6%	1.7%
R	.040	.035	.029	3.9%	3.5%	2.5%
PT	.139	.135	.244	**13.8%**	**13.8%**	**21.2%**
PR	.018	.017	.014	1.7%	1.8%	1.2%
TR	.007	.006	.014	0.7%	0.6%	1.5%
PTR	.147	.180	.190	14.6%	18.3%	16.5%

Note. Variance components pooled from eight analyses. D = delivery, L = language use, T = topic development.

D Studies

Changes in phi coefficients. In the G study, the variance components for different sources of variation were estimated. Using these variance components, D studies were conducted where the levels of the facets, which in this case were the number of raters and tasks, were varied to examine their impact on the phi coefficients of the analytic scores.

Table 7 provides the phi coefficients for the analytic scores when different combinations of number of raters and tasks are used for a fully crossed p x T x R design. One obvious observation was that the phi coefficients increase when more raters and more tasks are used. When one rating is obtained for each response using two or three tasks yields much higher phi coefficients than with one task, whereas when the number of tasks increases from four to six, the improvements in phi coefficients are much less dramatic, indicating a diminishing return when the number of tasks increases beyond four. The impact of increasing the number of ratings per response from one to two is modest and is mostly reflected in the reduction of the p x t x r and other undifferentiated error. Because the person-by-task interaction was relatively large compared to the effects concerning raters, increasing the number of tasks tends to have a larger impact on the phi coefficients.

The results of the D studies offer us useful information about optimizing assessment designs. On the one hand, the dependability and validity of the assessment need to be assured. On the other hand, cost for test development and scoring and efficiency of an assessment need to be factored in when designing an assessment. For example, it is worth noting that the phi coefficients when four tasks and one rater are used are expected to be higher than when two tasks and two raters are used. The same pattern is observed for six tasks and one rater versus three tasks and two raters.

The contrasting designs require the same total number of ratings; however, the single-rating-more-tasks design promises more dependable scores and at the same time ensures better domain coverage, if tasks are sampled appropriately. Therefore, if the cost for task development is less than obtaining multiple ratings per response, the single rating designs are more preferable given their cost-effectiveness.[4]

Changes in standard error of measurement. SEM provides some overlapping but also different information about the precision of examinees' scores than G and phi coefficients. It indicates on average the degree of uncertainty in a typical examinee's score (i.e., the difference between the observed and the universe scores of a typical examinee). One can then determine whether a certain amount of error is acceptable in practice given the purpose and the stakes of the assessment.

Table 7

Changes in Phi Coefficients of the Three Analytic Scores in D Studies

		Alternative D studies for p x T x R design					
		Single rating			Double rating		
	# of tasks	D	L	T	D	L	T
Phi coefficient	1	.65	.61	.55	.73	.70	.62
	2	.76	.74	.70	**.83**	**.81**	**.76**
	3	.81	.79	.76	***.87***	***.85***	***.82***
	4	**.83**	**.82**	**.80**	.89	.88	.85
	5	.85	.84	.82	.90	.89	.87
	6	***.86***	***.85***	***.84***	.91	.90	.89

Figures 1-3 plot the changes in the absolute-error SEMs in delivery, language use, and topic development scores when different combinations of number of raters and tasks are used. Comparing the three graphs shows that the absolute-error SEMs for topic development scores were higher than those for delivery and language use scores, due to the fact that the p x t interaction was the largest for topic development scores shown in Table 6. The differences were bigger when one or two tasks were used but dwindled when three or more tasks were used.

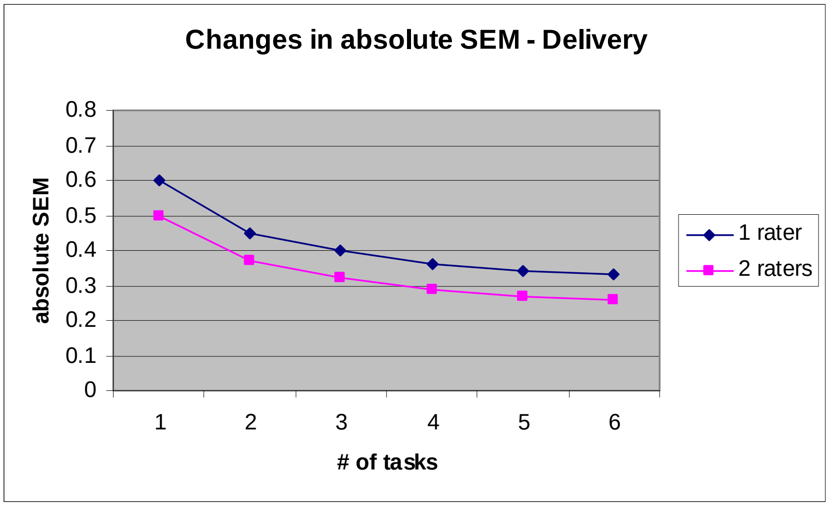

Figure 1. **Changes in absolute-error SEM for delivery scores.**

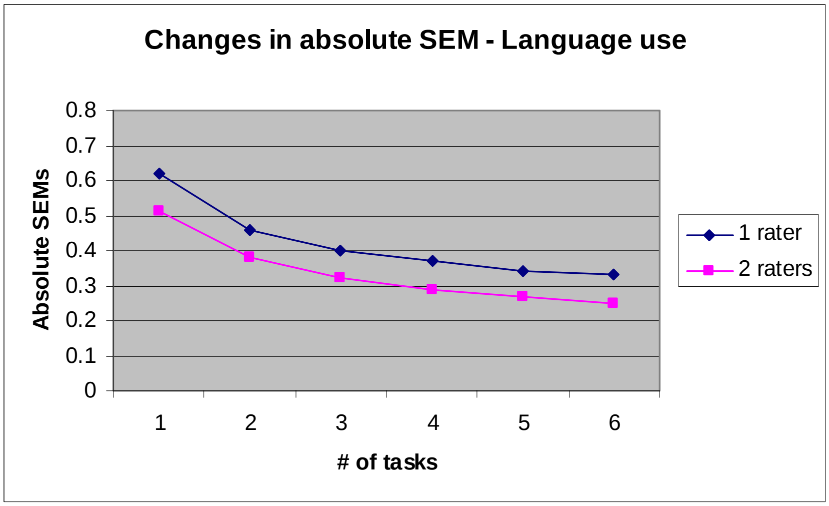

Figure 2. **Changes in absolute-error SEM for language use scores.**

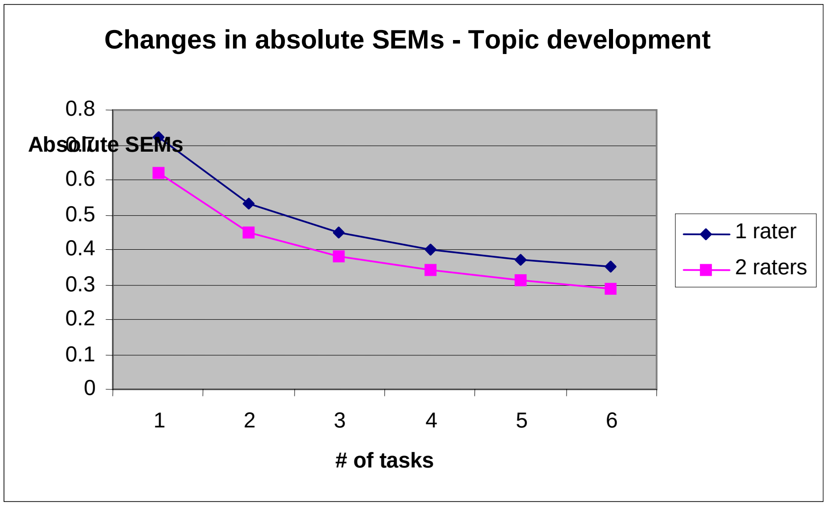

Figure 3. **Changes in absolute-error SEM for topic development scores.**

When six tasks and two raters are used, the absolute-error SEM for topic development scores is expected to be .29 on a scale of 1-4. When only two tasks and two ratings per task are obtained, the absolute-error SEM is expected to be .45 for topic development, which translates into a range of 1.8 points with a 95% confidence interval on a 1-4 point scale. This suggests a large error, given that there are only four score points.

As shown in Table 7, the phi coefficients are .91, .90, and .89 for delivery, language use, and topic development scores respectively when six tasks and two raters are used. These dependability estimates are considered as reasonably high for large-scale performance assessments. However, if one looks at the absolute-error SEMs, especially for topic development scores, a 95% interval spans a range of 1.2 on a scale of 1-4, which presents a less optimistic picture than the dependability estimates.

One has to bear in mind that both dependability indices and SEM in G theory indicate the precision of a measurement for a typical person and do not provide information at the individual examinee level.

The relative-error SEMs were smaller than the absolute-error SEMs and should be used when the decision is to distinguish examinees based on their scores. They are not discussed in detail here.

Dependability of Analytic Scores by Task

The analyses above illustrate, on average, how the phi coefficients would change with different combinations of number of tasks and raters. The analyses were based on averaged variance components for all tasks. Information on the dependability of analytic scores at the task level would also be useful because it could provide information on which tasks may introduce more unreliability in scoring a particular dimension.

Similar to the analyses on the dependability of analytic scores discussed in the previous section, for each task variance components were averaged from four different and independent analyses,[5] each one a fully crossed p x r design. The variance components associated with different sources of variation for different tasks are shown in Appendix G. Table 8 shows the results of the D studies, where the phi coefficients were compared for single versus double ratings for a p x R design.

As is shown in the table, the phi coefficients for Tasks 3-6 (integrated tasks) were generally higher than those for Tasks 1-2 (independent tasks), which suggests that the raters rated the three dimensions more consistently for the integrated tasks. Task 6 was an exception in that phi coefficients for language use and topic development were lower than those for the other integrated tasks. A careful examination of the item and responses reveals some problems in how the concepts of different kinds of money were explained and marked in the audio stimulus material. This resulted in inconsistency in examinee responses, which in turn made it difficult for raters to consistently judge the quality of topic development.

Table 8

Changes in Phi Coefficients by Task

		Single rating			Double ratings		
		Delivery	Language use	Topic development	Delivery	Language use	Topic development
Phi co-efficient	Task 1	.79	.61	.72	.88	.76	.84
	Task 2	.66	.71	.62	.79	.83	.77
	Task 3	.77	.65	.72	**.87**	**.79**	**.84**
	Task 4	.78	.74	.83	**.88**	**.85**	**.91**
	Task 5	.74	.79	.77	**.85**	**.88**	**.87**
	Task 6	.80	.65	.64	**.89**	**.79**	**.78**

Multivariate G Analyses on the Dependability of Composite Scores

In this study, each examinee was rated on three dimensions on all tasks and thus had three universe scores. When the dependability of the analytic scores was estimated separately, only the variance components were used. However, both the variance and covariance components were used in estimating the dependability of the composite scores, which were averages or weighted averages of the three analytic scores.

There were seven $p^{\bullet}$ x $t^{\bullet}$ x r° designs in Phase 2 rating, each row in Table 3 representing one $p^{\bullet}$ x $t^{\bullet}$ x r° design. In each $p^{\bullet}$ x $t^{\bullet}$ x r° design, persons were crossed with tasks with raters. Both raters and tasks are considered as random facets. All tasks by all persons were rated on all three dimensions, but a different rater pair rated a different dimension. Therefore, persons and tasks are indicated by a filled circle, whereas raters is marked by an unfilled circle, which represents that the same rater pair did not rate all the three dimensions. The covariance components involving raters such as *r*, *pr*, *tr*, and *prt* were all zero in those seven designs since different pairs of raters rated different dimensions in each design. Phase 1 rating featured a fully crossed $p^{\bullet}$ x $t^{\bullet}$ x $r^{\bullet}$ design, with persons crossed with tasks crossed with raters. Further, all three tasks taken by the 30 examinees were rated by the same raters on all three dimensions, so the object of measurement (*p*) and the two facets (*t* and *r*) are all marked by a filled circle.

Variance and covariance matrices from these eight analyses were pooled together to form one variance and covariance matrix so that all effects could be estimated. In the Phase 1 design, for the effects involving raters, only the diagonal values (variances) were used since the covariances among the three analytic scores associated with *r*, *p r*, *tr*, and *p t r* effects were zero in the seven $p^{\bullet}$ x $t^{\bullet}$ x r° designs.

The standard errors of the averages of the estimated variance components for the three analytic scores are given in Appendix H. The averages of the estimated covariance components and their standard errors are provided in Appendix H for *p*, *t*, and *pt* effects separately.

The SEs of the averages were generally small compared to the covariance components, suggesting that the averages were fairly stable estimates of the covariances among the three analytic scores for the person effect, the task effect, and the person-by-task interaction.

The final variance covariance matrix is shown in Table 9. The covariance components for different effects can be interpreted relative to the corresponding variance components for delivery, language use, and topic development. The covariance components provide some additional information about how examinees' universe scores on delivery, language use, and topic development covaried. They also showed how different error components associated with delivery, language use, and topic development scores covaried.

The covariance component for persons was the covariance between persons' universe scores on these three dimensions. The relatively high covariance components (as compared to the variance components) for persons indicate that examinees with high universe scores on delivery tended to have high universe scores on language use and topic development. The upper diagonal values for persons were the correlations among the universe scores on delivery, language use, and topic development, which will be discussed in detail in the section on the distinctness of the three dimensions. The covariances for the *pt* effect for these three dimensions were relatively large compared to the corresponding variance components. This suggests that the across-task differences in persons' relative standings were relatively consistent for the three dimensions. For example, persons' rank orderings on language use were to some extent different across tasks, but a somewhat similar pattern of differences in persons' rank orderings across tasks on topic development or delivery was observed. Both the variance components and covariance components for tasks were very small and are not discussed here.

If analytic scores are available for each task, it is possible to generate a composite score, which is an average, or weighted average, of the three analytic scores. This composite score indicates the overall quality of an examinee's response to a specific task. However, the interpretation of the composite score hinges on how the three dimensions are weighted. The common practice in generating a composite score is to give equal weight to all three scores, but for substantive reasons, one may want to give some dimensions more weight than others. Expert weights, which give rise to the profile of measures, are construct and theory driven and are preferable to sets of weights which maximize generalizability of composite scores (Webb & Shavelson, 1981). However, one should also explore the impact of using expert weights versus other weighting schemes on the dependability of composite scores.

Table 9

Estimated Variance and Covariance Components Pooled From

Seven $p^{\bullet}$ x $t^{\bullet}$ x r° Designs and One $p^{\bullet}$ x $t^{\bullet}$ x $r^{\bullet}$ Design

Effect	D	L	T
P	.657	**1.001**	**.981**
	.630	.604	**.982**
	.635	.611	.637
T	.003		
	-.007	.006	
	.002	-.007	.019
R	.040		
		.035	
			.029
PT	.139		
	.090	.135	
	.109	.123	.244
PR	.018		
		.017	
			.014
TR	.007		
		.006	
			.017
PTR	.147		
		.180	
			.190

Note. Lower diagonal elements are covariances.

Upper diagonal elements are universe-score correlations.

In this study, two weighting schemes were compared in terms of the dependability of the resulting composite scores. The first weighting scheme applied equal weights to all three dimensions. The second one was based on expert judgments about what was the assessment focus for each task type, where delivery and language use were given more weight for the independent tasks, and topic development was weighted more than delivery and language use for the integrated tasks (Table 10). Similar to the analysis performed earlier, the averaged variance and covariance components across four independent $p^{\bullet}$ x r° designs were used in the G studies for individual tasks (see Appendix I for the variance and covariance components for individual tasks). Then in the subsequent D studies, unit weights and expert weights were applied to the analytic scores, and phi coefficients of the composite scores were estimated for single rating versus double ratings.

Table 10

Two Weighting Schemes: Unit Weights (UW) and Expert Weights (EW)

	Independent tasks			Integrated tasks		
	Delivery	Language use	Topic development	Delivery	Language use	Topic development
Unit weights	1	1	1	1	1	1
Expert weights	1.5	1.5	1	1	1	1.5

The phi coefficients of the composite scores are shown in Table 11 for each task when unit weights and expert weights were used. The coefficients were comparable, supporting the use of expert weights, since they are construct-driven.

Distinctness of the Analytic Scores

An important criterion for evaluating the utility of analytic scoring is the extent to which it provides information over and above what holistic scores can offer. If an examinee gets similar scores across the three dimensions, the descriptions for typical performance at each score band in the holistic rubrics can be used to provide verbal diagnostic information for examinees. Therefore, a central issue here is to what extent the analytic scores are separable and to what extent the analytic scores differ. If the analytic scores are highly correlated and examinees show equal performance on these three dimensions, holistic scoring may be preferable given its efficiency and lower cost. This section discusses the separability of the three dimensions from two perspectives, the correlations among the analytic scores and raters' perceptions of the distinctness of the dimensions.

Table 11

Changes in Composite Score Phi Coefficients: Unit Weights (UW) Versus Expert Weights (EW)

	Phi			
	1 rater		2 raters	
	UW	EW	UW	EW
Task 1	.90	.90	.95	.95
Task 2	.90	.90	.95	.95
Task 3	.91	.91	.95	.95
Task 4	.92	.93	.96	.96
Task 5	.92	.92	.96	.96
Task 6	.90	.89	.94	.94

Correlations

The covariance components for persons (*p*) show how persons' universe scores on delivery, language use, and topic development covaried with one another. High covariance components among the three analytic scores relative to the variance components indicate that examinees who had high delivery scores tended to have high scores on language use or topic development. Disattenuated correlations among the three analytic scores were estimated based on the covariance components for persons (*p*) and the variance components for persons (*p*) in the multivariate G theory framework.

Table 12 compares the observed and disattenuated correlations among the three analytic scores by task. Disattenuated correlations answer this question: had the measurement been perfectly reliable, what correlations would be seen? They tell whether two sets of measurement have low observed correlations because of measurement error or because they are really uncorrelated. As is shown in the table, the observed correlations among the three analytic scores ranged from moderate to high. After correction for score unreliability, closer relationships among the three analytic scores were observed, as indicated by the disattenuated correlations among them. The disattenuated correlations between delivery and language use were the highest. Those between delivery and topic development scores by task were generally lower compared to those between delivery and language use and slightly lower than those between language use and topic development.

Table 12

Correlations Among the Analytic Scores: Observed Versus Disattenuated

	Delivery vs. language use		Delivery vs. topic development		Language use vs. topic development	
	Observed	Disattenuated	Observed	Disattenuated	Observed	Disattenuated
Task 1	.84	.96	.84	.92	.84	.93
Task 2	.79	.89	.78	.85	.81	.94
Task 3	.78	.92	.80	.88	.72	.83
Task 4	.90	.98	.86	.93	.88	.97
Task 5	.84	.91	.83	.87	.84	.88
Task 6	.72	.99	.75	.88	.69	.90
Average	.81	.94	.81	.89	.80	.91
All tasks	.94	1.00	.95	.98	.93	.98

Note. The disattenuated correlations for all tasks come from the variances and covariances pooled from eight analyses in Table 9.

Rater Perceptions of the Overlap Among Dimensions

When asked about their perceptions of the overlap among the three dimensions, up to half felt that there was some overlap (Table 13). In addition, six of them indicated that there was much overlap between delivery and language use, followed by five for language use and topic development. Only two of the raters thought the same for delivery and topic development, suggesting that raters perceived these two dimensions as the most distinct. In citing why they thought there was less overlap between delivery and topic development, they pointed to the ease of “shutting out” the content when rating delivery.

Raters' overall perceptions of the overlap among the three dimensions were also corroborated by their reactions to the statements designed to elicit similar information. For example, their general reaction to Statement 16, "*When rating a speaker with a heavy accent, I penalized him/her on language use and topic development since I could not understand him/her well enough to make an evaluation of other features,*" was disagreement, reflected by an average rating of 2.2, with 1 indicating "Strongly disagree." This suggests that the raters thought they could evaluate language use and topic development somewhat independently of delivery. Moreover, they generally agreed that if a speaker had a strong accent, they listened a few times to rate his/her language use and topic development (Statements 21 and 39; average ratings were 4.4 and 4.5). Their reactions to Statements 32, 33, 37, 39, and 42 point to the same conclusion: the raters thought there was less overlap between delivery and topic development and that in some aspects, they did not have much difficulty in distinguishing delivery from language use.

Table 13

Level of Overlap Among the Three Dimensions Perceived by Raters (N = 14)

Level of overlap	Delivery and language use		Delivery and topic development		Language use and topic development	
	Frequency	%	Frequency	%	Frequency	%
Very distinct	0	0.0%	4	28.6%	1	7.1%
Some overlap	6	42.9%	7	50.0%	7	50.0%
Much overlap	6	42.9%	2	14.3%	5	35.7%
Almost impossible to distinguish	0	0.0%	0	0.0%	0	0.0%
No response	2	14.3%	1	7.1%	1	7.1%

However, in other aspects, overlap between delivery and language use was perceived. Specifically, automaticity seems to be reflected in both delivery and language use. It is included as a feature in the descriptors for language use and manifested as a fluency feature in delivery. The raters reported that they were likely to pay a lot of attention to how effectively the speaker could put words and phases together "on the fly" when rating language use, as indicated by an average rating of 4.1 on Statement 26. Their responses to Statement 29, "W*hen rating language use, I focused on the precision and complexity of the speaker's vocabulary and grammar and rarely thought about automaticity,*" were generally negative, as shown by a rating of 2.9.

The raters also felt that there was substantial overlap between language use and topic development. The use of cohesive devices is a sticking point: whether a cohesive device is used correctly can be thought as part of language use, but at the same time the use of cohesive devices obviously has an impact on topic development. Therefore, the raters' struggle with this was revealed by their positive reactions to Statement 28, "*When rating language use, I found it hard to evaluate the use of cohesive devices without thinking about topic development,*" (average rating of 3.8). A similar perception is reflected in their responses to Statement 40, "*When rating topic development, I considered the effectiveness of the speaker's language use,*" where their mean rating was 3.2. The rubric for topic development makes many references to "clear progression of ideas" and "cohesion and coherence," but effective language use facilitates the progression of ideas, thus the raters may have felt that this was a gray area where connecting ideas plays a role in both language use and topic development.

Profile Variability

The observed correlations among the analytic scores on individual tasks ranged from moderate to high (Table 12). Technically, even when two sets of scores are perfectly correlated, varied profiles are still likely to exist, for example, when examinees are rank ordered exactly the same by the two sets of scores but one set of scores is consistently higher than the other. This scenario is not very likely to occur, though, unless the students in the sample are very homogeneous in their language acquisition and learning backgrounds. However, in this case, the means of the two sets of scores should be quite different. In this study, the correlations among the analytic scores by tasks ranged from moderate to high, but varied profiles at the task level are possible.

Table 14 demonstrates the proportion of any two analytic scores differing by 1 or more and by 1.5 or more by task (Columns 2-7), by task types (Columns 9-13) and when averaged across all tasks (Column 14).

Overall, for about one third of cases (overall by task: 35.1%), any two analytic scores differed by 1 or more. About 11% of the cases had two analytic scores differing by 1.5 or more. Further examination of the data did not reveal any patterns of profile scores by native language groups in this sample. However, when the analytic scores were averaged across task types that use different speaking contexts (Tasks 1 & 2: Speaking about familiar topics; Tasks 3 & 5: Speaking about campus life; Tasks 4 & 6: Speaking about academic course content) or that engage different combinations of skills (Tasks 3 & 4: Reading-listening-speaking tasks; Tasks 5 & 6: Listening-speaking tasks), they became less varied and we saw almost no varied profiles when the analytic scores were averaged across all tasks. The bottom-line question is this: If 11% of the examinees show varied profiles on some of the tasks, is analytic scoring warranted in operational settings given its complexity and inefficiency?

Table 14

Proportion of Any Two Analytic Scores Differing by 1 or More and by 1.5 or More

	1 or 1+	1.5 or 1.5+
T1	33.6	10.0
T2	35.0	8.6
T3	42.9	11.4
T4	20.7	5.7
T5	32.1	12.9
T6	46.4	15.0
Overall by task	35.1	10.6
Avg. of T1 & T2	10.0	0
Avg. of T3 & T5	6.4	0
Avg. of T4 & T6	4.3	0
Avg. of T3 & T4	5.7	0
Avg. of T5 & T6	11.4	0
Avg. all tasks	1.0	0

Table 14 shows observed within-person profile variability. Since there were measurement errors in the observed analytic scores, observed within-person profile variability may not reflect true variability. One could compute an index *Ģ*, which is the proportion of true within-person profile variability to observed within-person profile variability (Brennan, 2001). The idea is to compare true within-person variability derived from universe score variances and covariances to observed within-person variability based on observed score variances and covariances. It is similar to a reliability index, showing the reliability of observed within-person variability.

In Table 15, the observed and true within-person variability and the Ģ index are shown for each task for each dimension pair. They were computed for all possible pairs of the three dimensions because within-person variability across any two dimensions would warrant consideration in reporting profile scores. The observed within-person variability can be viewed as a measure of the flatness of the profile of observed scores whereas true within-person variability indicates the flatness of the profile of universe scores. The Ģ index is interpreted as the proportion of the variance in the profile of observed scores explained by the variance in the profile of universe scores for a randomly selected person in the sample. For example, considering Task 5, for a typical person, 56% of the within-person variance in the observed delivery and language use scores was attributable to the within-person variance in the universe scores of delivery and language use.

The Ģ indices were generally low, especially those for Task 4 and Task 1. Those for Task 1 were higher than for Task 4, but they were still lower compared to the other tasks. This raises some concern about using the observed profiles of scores as the indicator of the flatness of the profiles. It should be noted that this index of relative profile variability shows the average results for a typical person, so not all information in the data was captured.

Relationship Between the Holistic Scores and the Analytic Scores

In holistic scoring, raters consider the combined impact of delivery, language use, and topic development and make a global judgment about a person's performance on a particular task. During this process, raters attempt to weigh the impact of different dimensions on the overall effectiveness of communication to come up with a holistic score. Although there may be some commonalities in their perceptions of which linguistic weaknesses impact the effectiveness of communication most negatively or which linguistic strengths have the most impact on overall communication, there could well be individual differences in their perceptions of how all these different components together impact the communication of the global message. In other words, they may weight different components differentially and assign different holistic scores to the same response, or they may assign the same score to a response for different reasons.

Table 15

Observed Versus True Variability Across the Analytic Scores

	Pairs	Observed within-person variability	True within-person variability	Relative variability ($\hat{\mathcal{G}}$)
Task 1	D&L	.083	.035	.43
	D&T	.089	.035	.39
	L&T	.095	.041	.43
Task 2	D&L	.106	.057	.54
	D&T	.115	.067	.58
	L&T	.080	.025	.32
Task 3	D&L	.071	.033	.46
	D&T	.089	.054	.54
	L&T	.115	.074	.65
Task 4	D&L	.058	.014	.23
	D&T	.052	.009	.18
	L&T	.065	.018	.27
Task 5	D&L	.084	.047	.56
	D&T	.105	.069	.65
	L&T	.088	.051	.58
Task 6	D&L	.061	.016	.26
	D&T	.088	.044	.50
	L&T	.098	.047	.48

Note. Number of ratings/task = 2.

Because all 140 examinees were rated both holistically and analytically, this study could examine the relationships between the holistic and the analytic scores, which will shed light on which analytic scores drive the holistic scores. Table 16 illustrates the observed and disattenuated correlations among the dimension and holistic scores.

An overall observation is that there was a strong relationship between analytic scores and holistic scores for both independent tasks (Tasks 1 and 2) and integrated tasks (Tasks 3–6), as shown by disattenuated correlations corrected for score unreliability. No noticeable differences in the strengths of relationships were observed across the independent and integrated tasks.

Table 16

Correlations Among Holistic and Analytic Scores: Observed Versus Disattenuated

	Holistic vs. delivery		Holistic vs. language use		Holistic vs. topic development	
	Observed	Disattenuated	Observed	Disattenuated	Observed	Disattenuated
Task 1	.85	.98	.87	1.00	.84	.98
Task 2	.72	.87	.72	.89	.75	.92
Task 3	.79	.90	.75	.85	.85	.96
Task 4	.82	.91	.86	.97	.86	.96
Task 5	.87	.96	.86	.96	.96	1.00
Task 6	.77	.89	.80	.95	.77	.89

Note. The G coefficients with double ratings for both holistic scores and analytic scores were used to compute the disattenuated correlations between them. Note that 79 of the 140 cases were double-rated holistically so the G coefficients for holistic scores based on a sample of 79 were actually overestimates of the reliability for the 140 sample, hence the disattenuated correlations were underestimates.

Related to this, if this study used the holistic scores derived from overall impressionistic judgment and the composite scores, which are averages or weighted averages of the three analytic scores, how would the relative standings of examinees change? The correlations between the holistic scores and the composite scores derived from both unit weights and expert weights (see Table 10) are provided in Table 17.

This table shows that when averages or weighted averages of the analytic scores are used to rank order students as compared to the holistic scores, the relative standings of examinees would be similar. The rank orderings of examinees would be very similar irrespective of whether holistic scores of individual tasks or composite scores of individual tasks, based on unit weights or expert weights, were used.

Table 17

Correlations Between Holistic Scores and Composite Scores Based on Unit Weights and Expert Weights

	Holistic vs. composite score (unit weights)	Holistic vs. composite score (expert weights)
Task 1	.90	.90
Task 2	.79	.79
Task 3	.87	.88
Task 4	.88	.89
Task 5	.91	.91
Task 6	.87	.86
Sum of all task scores	.92	.92

Discussion

This section provides interpretations of the results to answer the five research questions.

Question 1: To What Extent Are the Analytic Scores Dependable?

The results from the rater agreement analysis and the G studies provide answers to this question.

Rater agreement. Overall, there was relatively little difference in the combined exact and adjacent agreement rates across the three dimensions (94.5% to 97.1%). Generally, the high nonadjacent agreement rates (2.9%–5.5%) would cause concern. However, this was the first time that the raters were trained to rate using the analytic rubrics. With more familiarity and enhanced rater training, their ratings are likely to become more consistent. The reasons reported by the raters in the survey for being confident or not confident about rating a particular dimension can provide valuable information about how rater training can be improved.

One of the two tasks with the highest nonadjacent disagreement rates on topic development scores was “problematic” in that the delineation of the two major concepts in the stimulus material was not clear. Thus the raters may have felt ambiguous about assessing the coverage and accuracy of major points. This calls for enhanced specifications for the organization and structure of the stimulus materials used for the integrated speaking tasks. Also required will be tighter topic notes for scoring topic development for each task that spell out both the main points that need to be covered and the penalties if less than complete or inaccurate information is presented.

Regarding the high nonadjacent agreement rates in delivery on some tasks, we have identified an isolated case where a rater who is familiar with the phonological patterns of a certain foreign language but does not have ESL/EFL teaching experience was more tolerant of accented speakers of this language and inclined to be more lenient in rating their delivery features. Studies in the second language acquisition literature have shown that familiarity with a particular accent facilitates intelligibility (Derwing & Munro, 1997; Gass & Varonis, 1984); however, it remains unclear how the judgments of speech quality by trained raters with or without ESL/EFL teaching experience may be affected by their familiarity with certain accents. This isolated case of inconsistency in rating delivery, if supported by further empirical evidence, might suggest that high tolerance of particular accents may be more of a problem for raters with no experience working with ESL/EFL students, whereas raters with ESL/EFL background may have developed an ear for distinguishing differing degrees of accents and be more likely to pose themselves as “naïve” listeners. However, the data did not allow us to observe a consistent pattern since the matching of rater background (familiarity with certain native languages and ESL/EFL experience) and examinee background (native language) was not systematically manipulated in this study. This line of research certainly deserves more attention as it has implications for whether it is justifiable to use raters who do not have ESL/EFL background to rate ESL/EFL speaking tests, which impact high-stakes decisions.

G studies. With regard to score dependability, the phi coefficients of the analytic scores for one task with a single rating were fairly low (.55–.65), but they improved somewhat with double ratings (.62–.73). When the analytic scores were averaged across two tasks and double ratings for each task, the phi coefficients rose to a reasonably high level (.76–.83). With six tasks and double ratings, they approached .90. This suggests that at the individual task level, when double ratings are obtained, the analytic scores would not be reliable enough for operational use but acceptable for low-stakes practice settings. With double ratings, the dependability of averaged analytic scores at the task type level (averages of two or more tasks) would be high enough for both operational and practice settings.

However, the SEMs suggest a less optimistic picture: when two tasks and two ratings per task were obtained, the absolute SEM for topic development scores was expected to be .45, translating into a span of 1.8 points with a 95% interval on a 0-4 point scale. This was a large error given that there are only four possible score points. When six tasks and double ratings were used, the span for topic development scores was 1.2, which was less optimistic than what the phi coefficient (.89) suggests but acceptable for operational settings.

G studies on the unadjudicated scores show that discrepancies in rater leniency accounted for a certain amount of error (5.6%, 6.1%, and 4.2% of the total variance of delivery, language use, and topic development scores respectively). This certainly deserves some attention in rater training and monitoring. The raters were more consistent in their rank ordering of examinees, the person-by-rater interaction explaining 3.4%, 1.8%, and 1.8% of the variance of delivery, language use, and topic development scores respectively. This suggests that although raters tended to agree on whether an examinee was better or worse on a certain dimension in comparison with others, they showed more discrepancies in the absolute score levels they assigned to examinees. Enhanced rater training materials that help raters interpret the score descriptors and find the appropriate score levels are thus needed.

When it comes to individual tasks, the raters showed more consistency in rating the integrated tasks than the independent tasks, suggesting the need for additional attention to rater training and monitoring on the independent tasks.

One of the largest error sources identified was the person-by-task interaction (13.8%, 13.8%, and 21.2% of the total variance of delivery, language use, and topic development scores respectively), which suggests that examinees' relative standings were very different across tasks. This would pose a serious threat to the dependability of the analytic scores (i.e., the generalization of analytic scores beyond the set of tasks used in the test if insufficient number of tasks were used).

It has been well-documented in educational measurement that performance-based tests introduce more variation across tasks; that is to say, some tasks may be difficult for some groups of test takers but not for others (Brennan & Johnson, 1995; Dunbar, Koretz, & Hoover, 1991; Gao, Shavelson, & Baxter, 1994; Lane, Liu, Ankenmann, & Stone, 1996; Linn, 1993; Linn & Burton, 1994; Shavelson, Baxter, & Gao, 1993; van der Vleuten & Swanson, 1990; Welch, 1991). However, in the language assessment literature, varying results have been reported with regard to the magnitude of the person- by-task interaction. This was mainly due to differences in the characteristics of the tasks and the scoring criteria used. If an assessment uses tasks that are not very differentiated in task types and in the ways tasks are contextualized and uses scoring criteria that are more driven by components that are relatively stable across tasks, it is less likely to see variation in performance across tasks. However, in those studies that reported large person-by-task interaction (e.g., Brennan et al., 1995; Lee, 2005; Lee & Kantor, 2005), the tasks were richly contextualized and/or the scoring rubrics contained features that were more task-specific, thus the quality of these features is more likely to vary across tasks.

In this study, the large person-by-task interaction was probably also attributable to the nature of the tasks and the scoring rubrics. The tasks in the TAST have been designed to integrate different types of skills (independent speaking vs. listening and speaking vs. reading, listening, and speaking) and to engage speaking skills in different contexts (everyday familiar topics vs. campus life vs. academic course content), so more variation in performance across tasks due to different task formats or contexts was expected. The scoring rubrics, which emphasize underlying abilities while including task-specific assessment focuses, also contributed to variation in examinees' scores across tasks. These seemingly conflicting findings are actually consistent with current theoretical models of communicative competence, which claim that communicative competence is to some extent stable while recognizing that some components may be local and dependent on the contexts in which the interactions occur (Chalhoub-Deville, 2003).

Variation in rater judgment and in performance across tasks has posed challenges for designing performance-based language tests. The variability in rater judgments can be reduced by rigorous development of scoring rubrics and rater training and monitoring; however, variation in examinees' performance across tasks presents more complex design issues (Brennan, 2000; Linn & Burton, 1994).

There are two potential ways to reduce the variability due to tasks: one is to increase the number of tasks, and the other is to reduce the person-by-task interactions in ways that would not weaken domain representation. There is a limit on the number of performance-based tasks that can be used in large-scale assessments due to logistic and efficiency concerns. If reducing the person-by-task interactions by manipulating domain specification and task sampling is attempted, there is a fine balance between reducing variation in performance across tasks and optimizing domain representation. Kane and his colleagues make a distinction between "universe of generalization" and the "target domain" (Kane, 1992; Kane, Crooks, & Cohen, 1999). With regard to task generalizability, G theory answers the statistical question of the consistency of examinees’ performance over samples of tasks, and it addresses the issue of whether scores obtained on a sample of tasks can be generalized to the universe that contains tasks similar to those included in the assessment. In other words, it addresses the strength of the link from “observed score” to “universe score,” as expressed in the interpretative validity argument by Kane and his associates. G theory can by no means provide evidence for establishing the link between performance on a sample of tasks (“observed score”) and expected performance in the target domain (“target score”), unless there is ample evidence to support that the universe of generalization and the target domain are similar.

If “cloned” tasks that are slight modifications of one another are used, less variation across tasks and improved score reliability might be seen; however, the universe of generalization is likely to be narrower and domain representation is likely to be lessened. In other words, task variability may be underestimated and generalizability may be overestimated to result in the weakening of the link between “universe score” and “target score.”

This design challenge of reducing sampling variability due to tasks warrants at least three types of studies. One type is job/needs analysis research that helps us precisely define the target domain and identify the underlying skills needed. The second type includes rigorously designed studies that manipulate the key task features that influence examinees’ performance on speaking tasks. The third type of studies investigate the relationship between examinee characteristics and task features to reveal the causes for person-by-task interactions. Through empirical research and emerging theories that build on the empirical results, one can identify construct irrelevant variables to control for task characteristics and to reduce person-by-task interactions and identify variables to vary to maximize construct representation, based on a well-defined theory of the domain. A key issue here are to define the domain tightly to the extent that it reflects the key subdomains informed by a job/needs analysis and sample. Another key is to design tasks carefully to the extent that they are representative of the target language use domain and engage the essential abilities and processes involved in real-world language use activities. Specifying the domain tightly can increase task homogeneity and potentially increase task reliability. Standardizing task characteristics in a principled way may also reduce variability arising from individual differences not relevant to the construct and thus reduce the variation in performance across tasks.

The sampling of speaking tasks in the TAST has drawn heavily on both current theories of communicative language ability and needs analysis studies to represent the underlying skills and processes engaged in speaking activities in an academic environment (Butler et al., 2000; Douglas, 1997; Rosenfeld et al., 2001; Waters, 1996). However, the second and third types of studies are still needed on the TAST to better understand the nature of person-by-task interactions and to reduce performance differences across tasks, to the extent possible.

Analytic scoring rubrics used in this study contain task-specific language features such as content coverage and accuracy, which may have exacerbated the problem of sampling variability due to tasks. On similar TOEFL iBT speaking prototype tasks, Lee (2005) reported that 16-17% of the total holistic score variance was explained by the person-by-task interaction. In this study, 21% of the variance in the topic development scores was attributable to the person-by-task interaction. This suggests that task-specific features may not have exerted as big an influence in holistic scoring when considered globally with other more stable features. However, if these task-specific language features are also what are valued in real language use contexts, efforts should be made to include them in the assessment criteria rather than keep them out to achieve better generalizability. If no attempt is made to measure the additional language features in oral discourse that performance-based speaking tasks can elicit, much of the advantage that performance-based language tests can bring is lost.

It should be noted that increased task sampling variability on some analytic dimensions was observed in comparison with results from one study that examined the reliability of holistic scores on similar tasks (Lee, 2005). More empirical evidence is certainly needed to support our speculation that task-specific features may have less influence in holistic scoring than in analytic scoring.

Implications of large task-sampling variability for reporting analytic scores. Variation in performance across tasks bears on the very key issues of ways to report analytic scores and stability of diagnostic feedback based on those analytic scores. Although analytic scoring seems to be an appealing approach to providing diagnostic score reports, deciding how to report the analytic scores and the descriptive information that goes with them is a difficult task. Factors to consider include whether the diagnostic report should be provided for the whole test, for different task types, or for different tasks. Each task or task type in TOEFL iBT speaking represents a whole universe of similar tasks in the domain, so information at task or task type level would be valuable. However, the generalizability of scores at task or task type level may be weakened since fewer tasks are used. If the dimension scores are averaged across all tasks and provide diagnostic information at the test level, the reliability will improve but very important information at the task or task type level will be missing.

In addition, variation across the three dimensions for the whole test may disappear since some students may perform better on one dimension on some tasks but not on others, which was seen in this study. This may suggest that in an operational setting, diagnostic information gleaned from the analytic scores for different task types could be unsatisfactory in terms of its generalizability to assessments that contain similar tasks.

However, the reliability estimates of analytic scores for individual tasks or task types with double ratings would be acceptable for low-stakes practice settings. In addition, in practice settings, it may be possible to carefully standardize task characteristics within a particular task type and also use more tasks for a particular task type. This way, it is possible to provide diagnostic information that is both reliable and useful for learners.

Question 2: How Is the Dependability of Composite Scores Impacted if Different Weighting Schemes Are Used?

It is possible to compute a composite score for each task, which is an average or weighted average of the three analytic scores. The composite score indicates the overall quality of a particular task response. However, the magnitude, interpretation, and dependability of the composite scores depend on how the dimensions are weighted. This study compares the dependability of composite scores derived from unit weights and expert weights. It was found that the use of expert weights versus unit weights had very little impact on the dependability of task-level composite scores. The impact of weighting schemes on composite score dependability is influenced by a few factors: the variances and reliability of the analytic scores and the correlations among the analytic scores. The comparable universe score variances of the three analytic scores and the generally high universe score correlations among the analytic scores by tasks (see Table 13) made the phi coefficients less sensitive to choice of weights and emphases (Wang & Stanley, 1970). Since the expert weights were determined based on substantive considerations, they preferable over unit weights.

Question 3: To What Extent Are the Dimensions Separable?

The disattenuated correlations among the three analytic scores by task ranged from moderately high to high, although those between delivery and topic development were lower than those between delivery and language use and slightly lower than those between language use and topic development. The relative distinctness of delivery and topic development was likely due to two factors: first, conceptually there is little overlap between delivery and topic development; second, operationally the descriptors for delivery and topic development were very distinct and the raters also thought it was easier to rate delivery in isolation from content, as revealed by the survey results.

A close examination of the rubrics reveals that the overlap in the descriptors for delivery and language use may partially explain the relatively high disattenuated-by-task correlations between delivery and language use scores. Along with range, complexity, and precision of vocabulary and grammar, automaticity is considered a key aspect among the language use descriptors. One of the manifestations of automaticity in language use is lexical and grammatical fluency (i.e., the effort with which lexical items and syntactic forms are retrieved and processed). Therefore, raters’ assessment of automaticity may be confounded with the fluency features measured in delivery. This overlap between the rubrics was perceived by the raters, as indicated by the rater survey results.

The raters also thought that it was more difficult to separate language use and topic development because examinees with a better command of vocabulary and grammar were more likely to do a superior job elaborating and developing their ideas within the response time specified for each task. Hence, very high correlations between language use and topic development scores were observed.

A major finding in this study is that, overall, the disattenuated correlations among the three analytic scores averaged across all tasks were very high. This suggests that (a) the constructs underlying the three analytic scores were highly correlated; (b) the three dimensions may be distinct conceptually, but raters were unable to consistently interpret the descriptors for each dimension either because they could not distinguish among them or because there was overlap among the descriptors in the three dimensions; or (c) the three dimensions were distinct, the rubrics were well-defined and the raters were able to pull the dimensions apart; however, this sample did not include distinct learner groups, so similarities in their English learning practices and exposure to English may have lead to high correlations among their analytic scores.

Unfortunately, this study did not provide sufficient information for us to evaluate all these hypotheses. Some overlap between fluency features in delivery and automaticity in language use in the rubrics seems possible. The rater survey results offered another piece of evidence: that the raters found overlap among the dimensions, especially between language use and topic development. However, further evidence is needed to support the diversity of the sample and raters' ability to separate these three dimensions.

As is the case with any kind of theory building efforts, great care should be taken to make sure that sample selection is driven by hypotheses. Often times the selected samples are convenient samples due to various practical constraints. In these cases, the investigators should draw the reader's attention to the limitations of the sample and alert them to the limited generalizations of the inferences drawn based on the sample data.

A caveat with the data in this study was that although it used a stratified sample, it was small and may not have included distinct learner groups. As discussed earlier, the sample used in this study was selected to represent a typical TOEFL test taking population with varying proficiency levels and native language backgrounds to inform the development of the holistic score rubrics, but all of the participants were recruited in the United States and had some exposure to English in an English-speaking country. Although native languages may provide some indication of their profiles of English ability, their exposure to English and their language learning environments (e.g., age of arrival and length of residence in an English–speaking country) may have a larger impact on their strengths and weaknesses. Since analytic scoring was not planned from the very beginning, information about specific language learning backgrounds was not collected. It may well be that this sample did not include very distinct learner groups, in which case it may have been more likely to be composed of distinct profiles. For example, certain speakers may repeatedly not get top scores on speaking tests because of serious delivery problems (such as strong accent and intonation patterns heavily influenced by their L1) alone. Often these are speakers of varieties of English, such as Indian English or Singaporean English, who have had very limited exposure to standard varieties of English. On the other hand, we may have seen highly communicative students function very successfully in an English-speaking environment, but they may repeatedly not get top scores because of their fossilized errors.

However, examples of these two classes of students seemed to be few and far between in this sample. Remaining students fell into this varied mix of participants, and separating the scores seemed to be more related to the task than to ability.

A line for future research would be to replicate this study with a sample that is more diverse in language learning backgrounds and amount and type of exposure to English. These factors are hypothesized to influence students' development patterns of speaking skills, and therefore an examination of the extent to which the three dimensions are distinct or related may be more fruitful with a more diverse sample. Some carefully planned empirical studies will provide more compelling evidence for theory building.

Another line of research related to investigations of the distinctness of the dimensions is to employ different kinds of qualitative methodologies such as interviews and verbal protocols to look into the processes raters use to make judgments on examinees' performance on the three dimensions. In this study a survey was used so raters could reflect on their analytic scoring experience. Through the survey some evidence was collected to support how raters thought they were able to separate the three dimensions. More research is needed to provide more evidence about the extent to which raters can distinguish among them.

Of course, the distinctness of the dimensions would also hinge on the analytic rubrics used in a particular study, in addition to the sampling and rater issues discussed earlier. Each of the "analytic" dimensions used in this study incorporates a wide range of speech elements. For example, delivery includes pronunciation, intonation, and fluidity, and language use covers grammar and vocabulary. However, an examinee's performance on grammar and vocabulary, or pronunciation and fluidity, will not necessarily be the same. A finer-grained analytic scale might have yielded different conclusions.

In addition, some dimensions or constructs are psychologically distinct, but psychometrically they may not be distinct due to the similar language learning backgrounds of students in the sample. To make the diagnostic information based on analytic scores maximally useful, it might be necessary to consider the characteristics of the target examinee population. Given their developmental patterns and language learning and instruction backgrounds, an exploration of which constructs may be closely correlated and on which dimensions they are likely to demonstrate varied performance compared with the other dimensions would be necessary. For example, if the target population includes a large proportion of highly communicative students who can function very successfully in an English-speaking environment but their speech may be characterized by fossilized errors, one might want to include grammatical accuracy as one dimension so as to provide the target examinees with the most useful diagnostic information. However, there is a limit to the grain size of the features human raters are to evaluate, and this is also constrained by the practicality and efficiency large-scale testing programs require.

Question 4: To What Extent Are Examinees' Profiles Varied?

In this study, some varied profiles of analytic scores were observed at the individual task level, but the profiles were generally flat at the task type (two or more tasks) or test levels. In addition, the relative variability of the score profiles at the task level was generally low, casting doubts on using the observed profiles of scores to support the true within-person score variability. This suggests that although analytic scores could provide valuable information about candidates' strengths and weaknesses at the task level, this diagnostic information may be somewhat limited since it may have been different if a different task was used. When analytic scores are averaged across tasks, the varied profiles at the task level become flattened, suggesting that the descriptors for the three dimensions in the holistic rubrics would work just fine to provide more detailed feedback on examinees' performance on different dimensions.

The reasons discussed in the answer to Question 3 about the lack of distinctness of the dimensions, such as the diversity of the sample, the nature of the analytic rubrics, and the quality of the analytic ratings, may also explain the flatness of the score profiles in this study.. Furthermore, the possible score points of the analytic scoring rubrics (1-4) and the holistic scoring guidelines that raters were instructed to follow may partially explain the lack of variability in analytic scores found in this study. The scoring guidelines state that an examinee has to be on target for all three dimensions (4s on all dimensions) to receive a holistic score of 4 and has to be on target for at least two of the three dimensions to get a holistic score of 1, 2, or 3. This suggests that differences of 1 or more would only occur with examinees scoring 2 or 3 holistically if the raters in holistic scoring were using the holistic rubrics appropriately and consistently. In addition, the only scenarios where the difference would be greater than 1 would be where an examinee is "above" the target level on some dimensions while performing "below" target on others.

Question 5: What Is the Relationship Between the Holistic Scores and the Analytic Scores? Is the Relationship Similar for Independent and Integrated Tasks?

Overall, strong relationships existed between the holistic and analytic scores as shown by the highly disattenuated correlations between them. In addition, the strengths of relationships remained stable across independent and integrated tasks, suggesting that the three dimensions may have played similar roles in raters' holistic judgments of the quality of responses to both types of tasks. A related finding was that if one were to use holistic scores or composite scores derived from averages or weighted averages of the analytic scores, examinees would be rank ordered very similarly based on their individual task scores (r =.79 -.91) or their whole test scores (r = .92).

Conclusion

This study investigated mainly the utility of analytic scoring for a large-scale speaking assessment by looking into the dependability of analytic scores and composite scores, the distinctness of the dimensions based on their correlations and raters' perceptions, and the flatness of examinees' profiles of analytic scores.

There was relatively little difference in the combined exact and adjacent rater agreement rates across the three dimensions, yet the high nonadjacent agreement rates cause concern. G studies show that the phi coefficients of the three analytic scores for one task with a single rating were fairly low (.55–.65), were reasonable for two tasks with double ratings (.76–.83), and approached .90 with six tasks and double ratings. However, the SEMs suggest a less optimistic picture for two-task double-rating scenarios. It was suggested that analytic scores averaged across two tasks and double ratings per task may not be reliable enough for operational use but acceptable for practice settings. Scores averaged across six tasks and double ratings would yield high reliability estimates appropriate for both operational and practice settings.

It was found that while raters tended to agree on the relative standing of examinees on a certain dimension, they showed more discrepancies in the absolute scores they assigned to examinees. The raters were also more consistent in rating the integrated tasks analytically than in rating the independent tasks. In addition, the person-by-task interaction contributed more to the variation in examinees' analytic scores than raters.

Disattenuated correlations among the analytic scores by task were high but those between delivery and topic development were generally lower. These results were corroborated by raters' perceptions. When averaged across tasks, observed and disattenuated correlations among the analytic scores were very high, and profiles of analytic scores for individual examinees were flat.

The results of this study suggest that it is not time for analytic score reporting, nor would there be much gain by using analytic scoring in operational settings. This is because, based on the sample used in this study, if averaged across tasks, the three analytic scores would be highly correlated and the profiles of scores flat, so although the analytic scores would be very reliable they would not provide additional information beyond what the holistic scores could offer for most examinees. On the other hand, if analytic scores on different tasks or task types are reported where there tend to be more varied profiles of analytic scores, the reliability would be too low given the stakes of the test.

The conclusions in this study, although tentative, could offer some useful information for the TOEFL program to consider when making decisions about scoring services related to TOEFL iBT speaking. Further investigations with samples of distinct learner groups, different analytic rubrics, and raters who are more familiar with analytic scoring and better trained may suggest otherwise about the value of information provided by analytic scores for operational use. More research is needed that addresses the limitations of the present study to yield more conclusive results.

An implication of this study for operational holistic scoring for TOEFL iBT speaking is that examinees with varied profiles can be flagged for analytic scoring if one wants the descriptive score reports to be dependable and useful for all examinees. For example, if a rater feels that an examinee’s performance on any two dimensions may differ by more than one level, he or she can flag that case for further scrutiny.

Information about performance on individual dimensions would be very useful for practice and self-learning where the stakes are much lower. For example, it may also be useful to provide a training package on analytic scoring for the speaking section of the institutional TOEFL iBT so that institutions can use it for placement and diagnostic purposes. Or analytic scoring can be provided as an additional service for potential TOEFL examinees who want to practice their speaking skills and improve their performance on the speaking section.

The findings of this study may help improve the training of holistic scorers for TOEFL iBT speaking. Specifically, the information on the score profiles (patterns of analytic scores) for different holistic score levels may shed light on how raters’ perceptions of performance on the three dimensions impact their global judgment. The analytic scores in this study could also be used for checking whether the holistic scoring guidelines—a holistic score of 4 requires “on target” performance on all three dimensions whereas holistic scores of 1, 2 or 3 require “on target” performance on at least two of the three—were applied appropriately and consistently.. Further, the patterns of analytic scores for different holistic score categories may inform the descriptive score report for examinees in operational settings. These research directions are certainly worth pursuing to benefit holistic scoring training and score reporting for TOEFL iBT speaking.

References

Bachman, L. F. (1988). Problems in examining the validity of the ACTFL oral proficiency interview. *Studies in Second Language Acquisition, 10*(2), 149-164.

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL Oral Interview. *The Modern Language Journal, 70*(4), 380-390.

Brennan, R. L. (1992). *Elements of generalizability theory* (Rev. ed.). Iowa City, IA: American College Testing.

Brennan, R. L. (2000). (Mis)conceptions about generalizability theory. *Educational Measurement: Issues and Practice*, *19*(1), 5-10.

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

Brennan, R. L. (2002). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24*(4), 339-353.

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.

Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice, 14*(4), 9-12.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12,* 1-15.

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

Chiu, C. W. T. (1999). *Scoring performance assessments based on judgments: Utilizing meta-analysis to estimate variance components in generalizability theory for unbalanced situations.* Unpublished doctoral dissertation, Michigan State University, East Lansing.

Chiu, C. W. T. (2000, April). *A subdividing method for generalizability theory: Precision of measurement errors and patterns of missing data.* Paper presented at the annual meeting of the American Educational Research Association (AERA), New Orleans, LA.

Chiu, C. W. T., & Wolfe, E. W. (2002). A method for analyzing sparse data matrices in the generalizability theory framework. *Applied Psychological Measurement, 26*(3), 321-338.

Chalhoub-Deville, M. (2003). Second language interaction: current perspectives and future trends. *Language Testing, 20*(4), 369-383.

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

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 19,* 1-16.

Douglas, D. (1997). *Testing Speaking Ability in Academic Contexts: Theoretical considerations* (TOEFL Monograph Series No. MS-08). Princeton, NJ: ETS.

Douglas, D., & Smith, J. (1997). *Theoretical underpinnings of the Test of Spoken English revision project* (TOEFL Monograph Series No. 9). Princeton, NJ: ETS.

Dunbar, S. B., Koretz, D., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4,* 289-304.

ETS. (2004). *iBT/Next Generation TOEFL test independent and integrated speaking rubrics (scoring standards)*. Retrieved February 8, 2006, from http://www.ets.org/Media/Tests/TOEFL/pdf/Speaking_Rubrics.pdf

Fulcher, G. (1997). The testing of speaking in a second language. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education. Vol 7: Language testing and assessment* (pp. 75-85). New York: Springer-Verlag.

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*, 323-342.

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of non-native speech. *Language Learning, 34,* 65-89.

Ingram, D., & Wylie, E. (1993). Assessing speaking proficiency in the international English language testing system. In D. Douglas & C. Chapelle. (Eds.), *A new decade of language testing research: Selected papers from the 1990 Language Testing Research Colloquium.* Alexandria, VA: TESOL, Inc.

Joe, G. W., & Woodward, J. A. (1976). Some developments in multivariate generalizability. *Psychometrika*, *41*, 205-217.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527-535.

Kane, M. T., Crooks, T. J., & Cohen, A. S. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17.

Lane, S., Liu, M., Ankenmann, R. D., & Stone, C. A. (1996). Generalizability and validity of a mathematical performance assessment. *Journal of Educational Measurement, 33,* 71-92.

Lee, Y.-W. (2005). *Dependability of scores for a new ESL speaking test: Evaluating prototype tasks* (TOEFL Res. Monograph No. MS-28). Princeton, NJ: ETS.

Lee, Y.-W., & Kantor, R. (2005). *Dependability of new ESL Writing test scores: Evaluating prototype tasks and alternative rating schemes* (TOEFL Res. Monograph No. MS-31). Princeton, NJ: ETS.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15,* 1-16.

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice, 13*(1)*,* 5-15.

Marcoulides, G. A. (1994). Selecting weighting schemes in multivariate generalizability studies. *Educational and Psychological Measurement, 54*(91), 3-7.

Rosenfeld, M., Leung, S., & Oltman, P. K. (2001). *The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels* (TOEFL Monograph Series No. MS-21). Princeton, NJ: ETS.

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

Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology, 34*, 133-166.

Smith, P. L. (1980, April). *Some approaches to determining the stability of estimated variance components.* Paper presented at the annual meeting of the American National Research Association (AERA), Boston, MA.

Underhill, N. (1987). *Testing spoken English.* Cambridge, England: Cambridge University Press.

Van der Vleuten, C. P. M., & Swanson, D. B. (1990). Assessment of clinical skills with standardized patients: The state of the art. *Teaching and Learning in Medicine, 2,* 58-76.

Wang, M. D., & Stanley, J.C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research, 40,* 663-705.

Waters, A. (1996). *A review of research into needs in English for academic purposes of relevance to the North American higher education context* (TOEFL Monograph Series No. MS-6). Princeton, NJ: ETS.

Weir, C. J. (1990). *Communciaitve language testing*. Englewood Cliffs, NJ: Prentice Hall.

Welch, C. (1991, April). *Estimating the reliability of a direct measure of writing through generalizability theory.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Webb, N. M., & Shavelson, R. J. (1981). Multivariate generalizability of general educational development ratings. *Journal of Educational Measurement, 18,* 13-22.

Webb, N. M., Shavelson, R. J., & Maddahian, E. (1983). Multivariate generalizability theory. In L. J. Fyans, Jr. (Ed.), *Generalizability theory: Inferences and practical applications: Vol. 18. New directions for testing and measurement* (pp. 67-81). San Francisco: Josey-Bass.

Wiley, D. E. (1992). *Studies of the California Assessment Program spring 1992 field trials*: *II. Pooling variance component estimates from random effects analyses of variance used to evaluate the generalizability of test tasks and scores.* Retrieved February 3, 2006, from http://www.c-save.umd.edu/poolingvarancecomponentsing-studies.pdf

Notes

[1] The other 30 examinees were quadruple-rated on Tasks 2, 4, and 5 in Phase 1.

[2] Although the same training materials and trainers were used in Phase 1 and Phase 2, the training and rating for Phase 1 rating were done at a different time. Therefore, a separate analysis was done to pool variance components from the seven p x t x r analyses only. The resulting variance components were similar as they were obtained when variance components were pooled from these eight analyses.

[3] Since the analysis was based on adjudicated scores, effects involving raters were underestimated. A similar analysis was conducted on nonadjudicated scores. The rater main effects accounted for 5.6%, 6.1%, and 4.2%, and the p x r interaction accounted for 3.4%, 1.8%, and 1.8% of the variance in delivery, language use, and topic development scores respectively, slightly higher than when adjudicated scores were used. However, the p x t interactions (11.1%, 9.7%, and 15.9%) based on the nonadjudicated scores were still much larger than the rater main effect and the p x r interactions.

[4] It should be noted that adjudicated scores were used in the analyses, so the errors concerning raters were actually underestimated. This conclusion needs to be buttressed by analyses based on nonadjudicated scores, especially for single rating scenarios. Analyses conducted on the nonadjudicated scores showed that holding the total number of ratings constant, the single rating scenarios generally provided slightly higher, or at least comparable, phi coefficients compared to the double rating scenarios. This was the first time raters used the analytic scoring rubrics, so with rigorous rater training and raters’ increasing familiarity with the analytic scoring rubrics, rater errors are likely to decrease.

[5] Each of the four examinee blocks was rated by a different rater pair/block on each task. Therefore, the variance components for each task were pooled from four independent p x r analyses. They were independent because in each design a different rater pair and a different examinee block were used.

List of Appendixes

Appendix A

TOEFL Academic Speaking Test Scoring Rubric

Independent Tasks

Score	General description	Delivery	Language use	Topic development
4	The response fulfills the demands of the task, with at most minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse. A response at this level is characterized by all of the following:	Generally well-paced flow (fluid expression). Speech is clear. It may include minor lapses, or minor difficulties with pronunciation or intonation patterns, which do not affect overall intelligibility.	The response demonstrates effective use of grammar and vocabulary. It exhibits a fairly high degree of automaticity with good control of basic and complex structures (as appropriate). Some minor (or systematic) errors are noticeable but do not obscure meaning.	Response is sustained and sufficient to the task. It is generally well developed and coherent; relationships between ideas are clear (or clear progression of ideas).
3	The response addresses the task appropriately, but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression, though it exhibits some noticeable lapses in the expression of ideas. A response at this level is characterized by at least two of the following:	Speech is generally clear, with some fluidity of expression, though minor difficulties with pronunciation, intonation, or pacing are noticeable and may require listener effort at times (though overall intelligibility is not significantly affected).	The response demonstrates fairly automatic and effective use of grammar and vocabulary, and fairly coherent expression of relevant ideas. Response may exhibit some imprecise or inaccurate use of vocabulary or grammatical structures or be somewhat limited in the range of structures used. This may affect overall fluency, but it does not seriously interfere with the communication of the message.	Response is mostly coherent and sustained and conveys relevant ideas/information. Overall development is somewhat limited, and usually lacks elaboration or specificity. Relationships between ideas may at times not be immediately clear.
2	The response addresses the task, but development of the topic is limited. It contains intelligible speech, although problems with delivery and/or overall coherence occur; meaning may be obscured in places. A response at this level is characterized by at least two of the following:	Speech is basically intelligible, though listener effort is needed because of unclear articulation, awkward intonation, or choppy rhythm/pace; meaning may be obscured in places.	The response demonstrates limited range and control of grammar and vocabulary. These limitations often prevent full expression of ideas. For the most part, only basic sentence structures are used successfully and spoken with fluidity. Structures and vocabulary may express mainly simple (short) and/or general propositions, with simple or unclear connections made among them (serial listing, conjunction, juxtaposition).	The response is connected to the task, though the number of ideas presented or the development of ideas is limited. Mostly basic ideas are expressed with limited elaboration (details and support). At times relevant substance may be vaguely expressed or repetitious. Connections of ideas may be unclear.
1	The response is very limited in content and/or coherence or is only minimally connected to the task, or speech is largely unintelligible. A response at this level is characterized by at least two of the following:	Consistent pronunciation, stress, and intonation difficulties cause considerable listener effort; delivery is choppy, fragmented, or telegraphic; there are frequent pauses and hesitations.	Range and control of grammar and vocabulary severely limits (or prevents) expression of ideas and connections among ideas. Some low level responses may rely heavily on practiced or formulaic expressions.	Limited relevant content is expressed. The response generally lacks substance beyond expression of very basic ideas. Speaker may be unable to sustain speech to complete task and may rely heavily on repetition of the prompt.

0 Speaker makes no attempt to respond OR response is unrelated to the topic.

Integrated Tasks

Score	General Description	Delivery	Language use	Topic Development
4	The response fulfills the demands of the task, with at most minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse. A response at this level is characterized by all of the following:	Speech is generally clear, fluid and sustained. It may include minor lapses or minor difficulties with pronunciation or intonation. Pace may vary at times as speaker attempts to recall information. Overall intelligibility remains high.	The response demonstrates good control of basic and complex grammatical structures that allow for coherent, efficient (automatic) expression of relevant ideas. Contains generally effective word choice. Though some minor (or systematic) errors or imprecise use may be noticeable, they do not require listener effort (or obscure meaning).	The response presents a clear progression of ideas and conveys the relevant information required by the task. It includes appropriate detail, though it may have minor errors or minor omissions.
3	The response addresses the task appropriately, but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression, though it exhibits some noticeable lapses in the expression of ideas. A response at this level is characterized by at least two of the following:	Speech is generally clear, with some fluidity of expression, but it exhibits minor difficulties with pronunciation, intonation or pacing and may require some listener effort at times. Overall intelligibility remains good, however.	The response demonstrates fairly automatic and effective use of grammar and vocabulary, and fairly coherent expression of relevant ideas. Response may exhibit some imprecise or inaccurate use of vocabulary or grammatical structures or be somewhat limited in the range of structures used. Such limitations do not seriously interfere with the communication of the message.	The response is sustained and conveys relevant information required by the task. However, it exhibits some incompleteness, inaccuracy, lack of specificity with respect to content, or choppiness in the progression of ideas.
2	The response is connected to the task, though it may be missing some relevant information or contain inaccuracies. It contains some intelligible speech, but at times problems with intelligibility and/or overall coherence may obscure meaning. A response at this level is characterized by at least two of the following:	Speech is clear at times, though it exhibits problems with pronunciation, intonation or pacing and so may require significant listener effort. Speech may not be sustained at a consistent level throughout. Problems with intelligibility may obscure meaning in places (but not throughout).	The response is limited in the range and control of vocabulary and grammar demonstrated (some complex structures may be used, but typically contain errors). This results in limited or vague expression of relevant ideas and imprecise or inaccurate connections. Automaticity of expression may only be evident at the phrasal level.	The response conveys some relevant information but is clearly incomplete or inaccurate. It is incomplete if it omits key ideas, makes vague reference to key ideas, or demonstrates limited development of important information. An inaccurate response demonstrates misunderstanding of key ideas from the stimulus. Typically, ideas expressed may not be well connected or cohesive so that familiarity with the stimulus is necessary in order to follow what is being discussed.
1	The response is very limited in content or coherence or is only minimally connected to the task. Speech may be largely unintelligible. A response at this level is characterized by at least two of the following:	Consistent pronunciation and intonation problems cause considerable listener effort and frequently obscure meaning. Delivery is choppy, fragmented, or telegraphic. Speech contains frequent pauses and hesitations.	Range and control of grammar and vocabulary severely limits (or prevents) expression of ideas and connections among ideas. Some very low-level responses may rely on isolated words or short utterances to communicate ideas.	The response fails to provide much relevant content. Ideas that are expressed are often inaccurate, limited to vague utterances, or repetitions (including repetition of prompt).

0 Speaker makes no attempt to respond OR response is unrelated to the topic.

Appendix B

Participant Requirements for the TAST Field Study

For this study you are asked to recruit participants who are currently enrolled in colleges or graduate schools or who are preparing to attend colleges (Intensive English Programs). The sample should include a mix of participants with varying levels of English language proficiency, varied first language backgrounds, and varied educational experience.

Our general aim is to recruit a sample that is as similar as possible to those who normally take the TOEFL in your area. Please keep this in mind as you recruit.

We would like you to recruit students with varying levels of English language proficiency as follows:

- Low – International students in high-intermediate intensive English classes who have not been admitted to an English language institution of higher education and may not yet be ready for admission. (Generally Paper-Based TOEFL scores in the range of 430 - 477, Computer-Based TOEFL score in the range of 117 - 153).

- Low/medium – International students in higher-level intensive English classes who have not been admitted to an English language institution of higher education but are nearly ready for admission. (Generally Paper-Based TOEFL scores in the range of 480 - 537, Computer-Based TOEFL score in the range of 157 - 203).

- Medium – International students who have been admitted to an English language institution of higher education within the past year but are currently enrolled in one or more English language support classes and who are also taking one or more credit bearing courses. (Generally Paper-Based TOEFL in the range of 540 - 577, Computer-Based TOEFL score in the range of 207 - 233).

- High – International students who have been admitted to an English language institution of higher education within the past year but were not required to take any ESL courses because their level of English proficiency was deemed sufficient due to previous educational experience, a local institutional test, or high TOEFL scores. (Generally Paper-Based TOEFL > 580, Computer-Based TOEFL score > 237).

Within each group, equal numbers of undergraduate and graduate students or equal numbers of students preparing for undergraduate and graduate study in English language institutions are desirable.

Appendix C

Native Language Backgrounds Represented by the 140-Person Analytic Scoring Sample

	Frequency	Percent
Arabic	9	6.4
Bangla	2	1.4
Bengali	3	2.1
Chinese	29	20.7
Ewe	1	0.7
Farsi	2	1.4
Finnish	1	0.7
French	1	0.7
French/Swahili	1	0.7
German	7	5.0
Gujerati	1	0.7
Hindi	5	3.6
Indian	1	0.7
Indonesian	1	0.7
Italian	1	0.7
Japanese	13	9.3
Korean	14	10.0
Kuraistan	1	0.7
Malay	1	0.7
Mongolian	1	0.7
Nepalese	1	0.7
Pakistani	1	0.7
Polish	2	1.4
Portuguese	4	2.9
Russian	5	3.6
Slovakian	1	0.7
Spanish	14	10.0
Swedish	1	0.7
Telugu	1	0.7
Thai	6	4.3
Turkish	4	2.9
Ukrainian	1	0.7
Venezuela	2	1.4
Vietnamese	2	1.4
Total	140	100.0

Appendix D

Descriptives of the Holistic Scores of the Analytic Scoring Sample

Table D1

Descriptives of the Holistic Ratings of the Analytic Scoring Sample

	Mean	SD
Item 1	2.64	1.02
Item 2	2.43	1.02
Item 3	2.57	0.96
Item 4	2.34	1.01
Item 5	2.44	1.05
Item 6	2.73	0.88
Total	2.52	0.86

Note. n = 140.

Table D2

Frequency Distribution of First Holistic Ratings for the Analytic Scoring Sample

Score level	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
0	7	4	1	3	2	0
1	10	26	22	26	29	12
2	**43**	**50**	**43**	**52**	**42**	**47**
3	**53**	**36**	**47**	**37**	**37**	**50**
4	27	24	27	22	30	31

Note. n = 140.

Appendix E

TAST Rater Questionnaire

Thank you for participating in the TAST analytic scoring study. Your valuable input during the training sessions will help us refine both the holistic and the analytic rating scales. As a follow-up, we would like you to provide some general background information and to reflect on your scoring experience. Please use "x" to indicate your choices.

I. General Background

1. Rater ID	

2. Native language	

3. What are your undergraduate/graduate specialties?	
Level	Specialty (ies)
Doctoral	
Master's	
Bachelor's	
Certificates	

4. Have you had any EFL/ ESL teaching experiences?	Yes	No
In which country	For how long	Courses you taught

5. Have you had any experience teaching language arts and English literature to English-as-L1 students?	Yes	No
If yes, for how long?		

6. What languages **other than English** do you understand, speak, read or write? Please list them below and indicate your level for each of the four skills (**use "1" for beginning, "2" for intermediate, and "3" for advanced.**)

Languages	***Understand***	***Speak***	***Read***	***Write***

7. What are the major native languages of the EFL/ESL students you have **most often** worked with?

8. What are the major native languages of the people that you have **often** had contact with (family members, colleagues, friends, etc.)?

9. How **difficult** is it for you to **understand** heavily accented speakers whose native languages you have studied or have been exposed to (the ones you listed in questions 6, 7, and 8)?				
Languages	**Very easy**			**Very difficult**
	1	**2**	**3**	**4**

10. Have you been a rater of other English oral test programs?	Yes	No

11. If yes, please complete the table below.

	Years
TSE	
SPEAK	
OPI	
Other (please specify)	

II. Reflections on the TAST Analytic Scoring

12. How would you rate your confidence level in rating delivery, language use, and topic development?				
	Not at all confident			Very
	confident			
	1	**2**	**3**	**4**
D				
L				
T				

13. Why did you feel confident or not confident about rating delivery, language use, or topic development?	
D	
L	
T	

14. Thinking back on your TAST analytic scoring experience, how much **overlap** did you find among these three dimensions?

	Very distinct	**Some overlap**	**Much overlap**	**Almost impossible to distinguish**
D and L				
D and T				
L and T				

15. Please rate the three dimensions in terms of how **important** you think they are in the communicative use of English for academic purposes.

	Somewhat Important			Extremely Important
	1	**2**	**3**	**4**
Delivery				
Language use				
Topic development				

Please choose the number that indicates the extent to which you agree or disagree with each of the following statements.

	Strongly Disagree					**Strongly Agree**
	1	**2**	**3**	**4**	**5**	**6**
Example: The training was excellent!						x
16. When rating a speaker with a heavy accent, I penalized him/her on language use and topic development since I could not understand him/her well enough to make an evaluation of other features.						
17. When rating delivery, I listened to the response only once if I believed that I had paid enough attention.						
18. When rating delivery, I was more likely to be distracted by a speaker's nonnative-like intonation and stress patterns than his/her problems pronouncing individual sounds.						
19. When rating delivery, I found it hard to make a judgment if the speaker 's delivery was very deliberate but clear.						
20. When rating delivery, I found it hard to make a fair judgment if I was familiar with the phonology of the speaker's native language.						
21.When rating a speaker with a strong accent, I listened a few times to rate his/her language use.						

22. When a response was particularly short, I found it hard to evaluate delivery.						
	Strongly Disagree					**Strongly Agree**
	1	**2**	**3**	**4**	**5**	**6**
23. When rating language use, I ignored lexical or grammatical errors that did not have much impact on the meaning conveyed.						
24. When rating language use, I found myself counting the number of errors the speaker made.						
25. When rating language use, I distinguished errors that had an impact on meaning and those that did not and penalized them differentially.						
26. When rating language use, I paid a lot of attention to how effectively the speaker could put words and phrases together “on the fly.”						
27. When rating language use, I paid a lot of attention to how effectively the speaker tied his ideas together.						
28. When rating language use, I found it hard to evaluate the use of cohesive devices without thinking about topic development.						
29.When rating language use, I focused on the precision and complexity of the speaker’s vocabulary and grammar and rarely thought about automaticity.						
30. When rating language use, I paid more attention to precision than range or complexity.						
31. When rating language use, I found myself making a mental note of what types of grammatical structures the speaker used.						
32. When rating language use, a speaker’s native-like pronunciation and intonation patterns affected my judgment.						
33. When rating language use, I tried to transcribe the speech in my mind.						
	Strongly Disagree					**Strongly Agree**
	1	**2**	**3**	**4**	**5**	**6**
34. If a response was particularly short, I found it hard to evaluate language use.						

35. When rating language use, I listened to the speech more than once if I thought that the speaker’s weaknesses in other areas would affect my judgment.						
36. When rating topic development, I tried to fill the gaps by using my knowledge about the prompt and stimuli and the expected response to speculate what the speaker was trying to get across.						
37. When rating topic development, a speaker’s native-like pronunciation and intonation patterns affected my judgment.						
38. When rating topic development, I listened to the speech more than once if I thought that the speaker’s language use was inadequate.						
39. When rating a speaker with a strong accent, I listened to the speech a few times for topic development.						
40. When rating topic development, I considered the effectiveness of the speaker’s language use.						
41. When rating topic development, it was hard to decide how to penalize a speaker if he/she presented inaccurate information *but* showed strengths in other dimensions.						
42. When rating topic development, I tried to transcribe the speech in my mind.						
43. When a response was particularly short, I found it hard to evaluate topic development.						

If you have other comments you would like to share with us, please provide them in the space below.

About the analytic rating rubric:

About the training materials:

About the training procedures:

About the analytic scoring experience:

Other:

Appendix F

Estimated G Study Variance Components

Table F1

Estimated G Study Variance Components for Delivery

Block	Source						
	Persons *(p)*	Tasks *(t)*	Raters *(r)*	*Pt*	*pr*	*tr*	*Ptr, e*
1	0.292	0.004	0.011	0.109	0.007	0.000	0.219
2	0.655	0.000	0.012	0.100	0.011	0.004	0.131
3	0.645	0.012	0.021	0.187	0.006	0.009	0.103
4	0.521	0.000	0.003	0.229	0.021	0.015	0.185
5	0.649	0.000	0.034	0.080	0.029	0.015	0.174
6	0.875	0.003	0.059	0.140	0.066	0.000	0.098
7	0.960	0.000	0.013	0.140	0.000	0.010	0.143
8	0.660	0.002	0.166	0.129	0.001	0.000	0.125
Average[a]	**0.657**	**0.003**	**0.040**	**0.139**	**0.018**	**0.007**	**0.147**
SE[b]	**0.072**	**0.001**	**0.019**	**0.017**	**0.008**	**0.002**	**0.015**

[a]Average of the variance component estimates from eight analyses. [b]Standard error of the average.

Table F2

Estimated G Study Variance Components for Language Use

Block	Source						
	Persons *(p)*	Tasks *(t)*	Raters *(r)*	*Pt*	*pr*	*tr*	*Ptr, e*
1	0.295	0.018	0.013	0.208	0.052	0.000	0.195
2	0.578	0.000	0.001	0.134	0.000	0.010	0.207
3	0.669	0.009	0.005	0.099	0.027	0.004	0.127
4	0.564	0.008	0.167	0.149	0.000	0.000	0.171
5	0.573	0.006	0.036	0.084	0.035	0.022	0.193
6	0.742	0.000	0.000	0.149	0.000	0.013	0.182
7	0.784	0.000	0.000	0.099	0.024	0.001	0.188
8	0.624	0.003	0.055	0.159	0.000	0.000	0.175
Average[a]	**0.604**	**0.006**	**0.035**	**0.135**	**0.017**	**0.006**	**0.180**
SE[b]	**0.053**	**0.002**	**0.020**	**0.014**	**0.007**	**0.003**	**0.009**

[a]Average of the variance component estimates from eight analyses. [b]Standard error of the average.

Table F3

Estimated G Study Variance Components for Topic Development

Block	Source						
	Persons *(p)*	Tasks *(t)*	Raters *(r)*	*Pt*	*pr*	*tr*	*Ptr, e*
1	0.203	0.039	0.095	0.354	0.018	0.018	0.147
2	0.588	0.000	0.034	0.244	0.000	0.000	0.230
3	0.988	0.000	0.023	0.167	0.035	0.035	0.219
4	0.465	0.021	0.000	0.384	0.000	0.000	0.236
5	0.692	0.000	0.014	0.203	0.014	0.014	0.263
6	0.613	0.048	0.037	0.142	0.000	0.000	0.152
7	0.909	0.000	0.027	0.181	0.018	0.018	0.123
8	0.640	0.047	0.000	0.276	0.026	0.026	0.149
Average[a]	**0.637**	**0.019**	**0.029**	**0.244**	**0.014**	**0.014**	**0.190**
SE[b]	**0.087**	**0.008**	**0.011**	**0.031**	**0.005**	**0.005**	**0.019**

[a] Average of the variance component estimates from eight analyses. [b] Standard error of the average.

Appendix G

G Study Variance Components by Tasks

		Variance component			Percent of total variation		
		D	L	T	D	L	T
Task 1	P	0.738	0.677	0.888	76.4%	74.2%	77.1%
	R	0.039	0.047	0.017	4.1%	5.1%	1.5%
	PR	0.188	0.189	0.246	19.5%	20.7%	21.4%
Task 2	P	0.892	0.789	0.833	78.7%	75.6%	75.4%
	R	0.074	0.031	0.052	6.5%	2.9%	4.7%
	PR	0.168	0.223	0.219	14.8%	21.4%	19.8%
Task 3	P	0.657	0.779	0.878	79.6%	78.1%	78.7%
	R	0.011	0.068	0.062	1.4%	6.8%	5.6%
	PR	0.158	0.151	0.175	19.1%	15.1%	15.7%
Task 4	P	0.976	0.812	0.984	80.1%	78.1%	84.0%
	R	0.080	0.029	0.007	6.6%	2.8%	0.6%
	PR	0.162	0.199	0.180	13.3%	19.1%	15.4%
Task 5	P	0.899	0.788	0.946	80.9%	81.7%	80.5%
	R	0.067	0.026	0.081	6.1%	2.7%	6.9%
	PR	0.145	0.151	0.148	13.0%	15.6%	12.6%

Note. D = delivery, L = language use, T = topic development.

Appendix H

Estimated G Study Covariance Components for the $p^{\bullet}$ x $t^{\bullet}$ x r° Design

Effect	*P*			*T*			*pt*		
Block	D&L	D&T	L&T	D&L	D&T	L&T	D&L	D&T	L&T
1	0.313	0.235	0.250	-0.020	0.010	-0.026	0.121	0.168	0.191
2	0.596	0.603	0.563	-0.004	-0.004	-0.001	0.112	0.118	0.132
3	0.681	0.831	0.827	-0.020	0.009	-0.009	0.052	0.071	0.069
4	0.537	0.486	0.477	-0.003	-0.011	-0.028	0.176	0.219	0.169
5	0.609	0.647	0.629	-0.004	0.003	-0.002	0.072	0.097	0.105
6	0.794	0.729	0.671	0.003	0.015	0.006	0.087	0.074	0.098
7	0.870	0.917	0.828	0.001	-0.001	-0.002	0.061	0.082	0.068
8	0.641	0.635	0.640	-0.008	-0.004	0.006	0.041	0.038	0.151
Average[a]	0.630	0.635	0.611	-0.007	0.002	-0.007	0.090	0.109	0.123
SE[b]	0.059	0.074	0.067	0.003	0.003	0.005	0.016	0.021	0.016

Note. Covariances pooled from eight analyses. D = delivery, L = language use, T = topic development.

[a] Average of the variance component estimates from eight analyses. [b] Standard error of the average.

Appendix I

Pooled Variance and Covariance Components for Individual Tasks

		P			R	PR
Task 1	D	.73841	.68012	.74839	.03940	.18836
	L		.67715	.72083	.04693	.18860
	T			.88833	.01739	.24590
Task 2	D	.89151	.74526	.73156	.07396	.16779
	L	.74526	.78913	.76479	.03074	.22335
	T	.73156	.76479	.83253	.05207	.21888
Task 3	D	.65713	.65654	.67202	.01135	.15754
	L	.65654	.77865	.68376	.06815	.15078
	T	.67202	.68376	.87838	.06234	.17490
Task 4	D	.97642	.87183	.91447	.07999	.16224
	L	.87183	.81212	.86277	.02906	.19889
	T	.91447	.86277	.98402	.00661	.18044
Task 5	D	.89938	.76887	.80454	.06730	.14501
	L	.76887	.78820	.76355	.02575	.15053
	T	.80454	.76355	.94559	.08102	.14789
Task 6	D	.65211	.60543	.62402	.00789	.15303
	L	.60543	.57529	.60369	.03258	.21134
	T	.62402	.60369	.77422	.06054	.19732

Note. Variances and covariances pooled from four analyses.

Appendix J

Relative-Error and Absolute-Error SEMs by Dimension by Task

	Double ratings					
	Relative-error SEM			Absolute-error SEM		
	D	L	T	D	L	T
Task 1	.31	.31	.35	.34	.34	.36
Task 2	.29	.33	.33	.35	.36	.37
Task 3	.28	.27	.30	.29	.33	.34
Task 4	.28	.31	.30	.35	.34	.31
Task 5	.27	.27	.27	.33	.30	.34
Task 6	.28	.33	.31	.28	.35	.36

Note. D = delivery, L = language use, T = topic development.

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)

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

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