ETS TOEFL

Quality Beyond Measure.

TOEFL iBT® Research Report
TOEFL iBT–20

Developing Analytic Rating Guides for *TOEFL iBT*® Integrated Speaking Tasks

Joan Jamieson

Kornwipa Poonpon

July 2013

Developing Analytic Rating Guides for TOEFL iBT's Integrated Speaking Tasks

Joan Jamieson
Northern Arizona University, Flagstaff

Kornwipa Poonpon
Khon Kaen University, Thailand

RR-13-13

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

Research and development of a new type of scoring rubric for the integrated speaking tasks of *TOEFL iBT*® are described. These *analytic rating guides* could be helpful if tasks modeled after those in TOEFL iBT were used for formative assessment, a purpose which is different from TOEFL iBT's primary use for admission decisions. Two questions motivated the project: What can be done to make the criteria and standards for good performance clear not only to English language teachers and their students but also to novice raters? How can test-takers be guided in the steps necessary to improve performance? Previous research, quantitative results of linguistic features present in spoken responses, and qualitative themes of raters' judgments were analyzed. Salient features associated with performance were then synthesized to develop three analytic rating guides, each using a series of yes/no questions. The rating guides expanded the current scoring scales for the three dimensions of delivery, language use, and topic development, so that key features of increasingly proficient speaking performance were described in more detail. Suggestions for future validation studies are provided in terms of an iterative process of trials and revisions.

Key words: ability testing, English as a second language, formative assessment, grammar, oral communication, phonology, scoring rubrics, verbal reports

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. The test is now wholly owned and operated by ETS.

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, two-year colleges, and nonprofit educational exchange agencies.

❖ ❖ ❖

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 (2012-2013) members of the TOEFL Committee of Examiners are:

John M. Norris - Chair	Georgetown University
Maureen Burke	The University of Iowa
Yuko Goto Butler	University of Pennsylvania
Barbara Hoekje	Drexel University
Ari Huhta	University of Jyväskylä, Finland
Eunice Eunhee Jang	University of Toronto, Canada
James Purpura	Teachers College, Columbia University
John Read	The University of Auckland, New Zealand
Carsten Roever	The University of Melbourne, Australia
Steve Ross	University of Maryland
Norbert Schmitt	University of Nottingham, UK
Ling Shi	University of British Columbia, Canada

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

Funding for this project was received from the TOEFL Committee of Examiners, ETS. The authors express their thanks to the following individuals at Northern Arizona University who provided support: Don Carter, Matthew Minster, and Lawrence Walters for developing online systems for training, scoring, and verbal reports; Doug Biber for lexico-grammatical tagging; Okim Kang for phonological analyses; Kim McDonough and Roy St. Laurant for statistical advice; and Alex Boruki, Rachel Donelson, Don Miller, Ozgür Parlak, and Diana Sanchez for transcribing. Finally the authors are grateful to Carolyn Turner and other anonymous reviewers, whose comments were extremely helpful in revising the draft report; while trying to incorporate their suggestions, the authors, of course, are responsible for the final content.

Table of Contents

List of Tables

Page

List of Figures

Page

Overview

Official test score users confirmed their desire to use the new Internet-based *TOEFL*® test, the *TOEFL iBT*® test, for its primary purpose—informing admission decisions to English-speaking higher-education institutions—in focus groups, interviews, and surveys. Users wanted the new version of the test to include constructed-response tasks involving speaking and writing. Also, they thought that the TOEFL iBT score would be useful in guiding English-language instruction (Chapelle, Enright, & Jamieson, 2008; Taylor & Angelis, 2008).

The English language skill most in need of guidance was speaking. The inclusion of the speaking section in the TOEFL iBT test was seen as the greatest change in the new version of the test and as the area that would have positive washback on the English language teaching and learning community around the world (Butler, Eignor, Jones, McNamara, & Suomi, 2000; Mackenzie, 2005; Sakui, 2004; Shohamy, 2006; Wall & Horák, 2006, 2008). The TOEFL iBT test's new integrated speaking tasks posed meaningful performance-based test tasks, including academic content such as readings and lectures. Textbook publishers, alert to the instructional potential of content-based speaking tasks and the market potential of changes in test preparation courses, provided similar tasks along with the TOEFL iBT test's holistic speaking rubrics so that English language teachers could raise students' awareness of their abilities in relation to the TOEFL iBT scale (e.g., *NorthStar: Building Skills for the TOEFL iBT*; Solórzano, 2006). However, those existing holistic rubrics were designed for admission decisions, not for formative assessment.

Any holistic rubric designed for an admission purpose might be simply too broad to capture students' language development and might not provide teachers with adequate guidance for formative classroom assessment (Colby-Kelly & Turner, 2007; Fulcher, 1996b; Rea-Dickins, 2004; Upshur & Turner, 1995). How can a rubric designed for one test purpose be appropriate for another? After a test has been developed, the determination of how suitable that test is for its intended use has been foundational in validation (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; American Psychological Association, 1954; Bachman, 1990; Davies, Brown, Elder, Hill, Lumley, & McNamara, 1999). In extending the use of tests, there needs to be a balance between the benefits of increased information to different audiences on the one side and the drawbacks of compromises in test design and cautions in interpretations on the other side (Messick, 1980;

Miller, Linn, & Gronlund, 2009). Over 60 years ago, Gulliksen (1950) described the potential benefits and drawbacks of using an English proficiency test of writing ability to inform instructional practices. Among the benefits, the test "might result in considerable reorganization of teaching so that differential types of learning ability would be recognized by teaching practices" (p. 513). Gulliksen thought that such a test had the potential to be a *powerful incentive* to both students and teachers. This idea resonated with Messick years later when he related how the benefits of transparency and meaningfulness of general performance assessments could be ascribed to classroom use:

> If assessment itself is to be a worthwhile educational experience serving to motivate and direct learning, then the problem and tasks posed should be meaningful to the students. That is, not only should students know what is being assessed, but the criteria and standards that constitute good performance should be clear to them in terms of both how the performance is to be scored and what steps might be taken or what directions moved in to improve performance…a number of things can be done…to improve somewhat the transparency and meaningfulness of structured and semi-structured tasks. (Messick, 1994, p. 16)

The study reported here addressed an ancillary use for the TOEFL iBT test—formative assessment. By addressing one of the TOEFL program research agenda items, our idea was to modify the integrated speaking rubric so that it would be appropriate for this different use.

Informed by the development of the ACTFL rating scales (American Council of Teachers of Foreign Languages; e.g., Clark & Clifford, 1988), we decided to expand the scale. Advantages associated with analytic scoring of expanded scales were particularly appealing (e.g., Sawaki, 2007; Xi & Mollaun, 2006). Also, novice raters might be helped by training with an analytic rating guide so that they might better understand how to balance elements when assigning a holistic score. Referring back to earlier words, we wanted to create a powerful incentive to improve speaking instruction by making the criteria and standards for good performance clear to English language learners and their teachers, as well as to novice raters, and to guide instruction along the steps associated with better performance.

Addressing these goals, the original 0–4 point holistic scale was expanded to three analytic 0–7 point rating guides for each of the dimensions included in TOEFL iBT's speaking rubric (i.e., delivery, language use, and topic development). Modeled after the work of Upshur

and Turner (1995), these *rating guides* were made up of a series of yes/no questions about important criteria for good speaking performance. The questions were intended to guide the test-taker, the teacher, or the novice rater not only to a score, but to an understanding of the linguistic features that underlie that score. The rating guides were based on qualitative and quantitative approaches to scale expansion. The qualitative approach included think-aloud protocols of factors that affected raters' decision-making while they were scoring spoken responses. The quantitative approach included the frequency of linguistic features in spoken responses and raters' scores. This report describes the development of these new analytic rating guides.

Background

To ensure that our project reflected current thoughts in applied linguistics yet remained maximally informative for TOEFL iBT, we reviewed work in two areas: (a) practices in scale development and (b) the role of linguistic features in speaking performance.

Scale Development

An effective rating scale should be driven by a construct, be meaningful in given target language use situations, and be informed by analysis of actual spoken responses. In second language testing, these three approaches (i.e., theoretical, functional, and empirical) have been used to underpin the development of rating scales with different rationales and purposes. The approach used to develop the scale is operationalized, in part, by the method of scoring that is chosen. After a brief overview of approaches and scoring, the development of TOEFL iBT's integrated speaking scale rating scale is reviewed to ensure that the new analytic rating guides not only reflect its subconstructs and claims, but also address its difficulties.

One approach to development of rating scales is based on theoretical models of language reflecting a belief that by explicitly linking observable behaviors with underlying language abilities, test results will be transparent (Hudson, 2005). In this approach, the inference of main interest in a validity argument is between performance and the construct, namely *explanation* (Chapelle et al., 2008). Evidence for validity relies on applied linguistics theories that provide insight for hypotheses about expected relationships with other variables (Bachman, 1990; Bachman & Savignon, 1986; Canale & Swain, 1980; Cronbach & Meehl, 1955; Embretson, 1983; Fulcher, 1996a, 1997; North, 2000). The revised TSE® exam (Douglas & Smith, 1997) was an example of this approach in assessing speaking.

The functional approach to scale development was influenced by the language that test takers were expected to use in target situations. In the functional approach the inference of interest, namely *extrapolation*, links performance to contexts of use (Chapelle et al., 2008). Evidence for validity relies on the predictions of examinees' abilities to handle language in their real world jobs or contexts (Dandonoli & Henning, 1990; Lowe, 1986; Stansfield & Kenyon, 1992). Notable examples include the Foreign Service Institute scale that was used to assess the language needed by government personnel in diplomatic or military contexts and the ACTFL guidelines, designed to assess university students' ability to function effectively in real-life contexts (Breiner-Sanders, Lowe, Miles, & Swender, 2000; Clark & Clifford, 1988; Liskin-Gasparro, 1984).

Both of these approaches, when used alone, have critics. For example, theoretically based scales assume a priori constructs of language ability that may or may not be relevant to speaking performance in a given situation as they often define decontextualized ability (Hudson, 2005; North, 2000). Despite their tradition, the functionally based ACTFL scales were viewed skeptically in part because they were developed intuitively (Bachman & Savignon, 1986; Fulcher, 1996b, 1997; North, 1993, 2000; Savignon, 1985). In fact, a concern with both theoretically based and functionally based scale development approaches was related to their lack of empirical validation.

Rather than solely relying on applied linguistics' theories or relevant contexts, recently scale developers have argued for including empirical support as fundamental evidence to justify scores (Fulcher, 1997; Hudson, 2005; Knoch, 2007a; McNamara, 1997). Much attention has been drawn to test takers' actual language use as key evidence to developing rating scales. Fulcher believed that a *data-driven approach* was promising in rating scale development (1993, 1996a). He used a combination of qualitative and quantitative analyses of learners' spoken language to develop his empirically based fluency scale. Such data-driven scales are based on analysis of learners' language gathered from target language test situations and contexts and have been recognized for their practicality and authenticity (Fulcher, 1987, 1993, 1996a; North, 1993, 1995, 2000; North & Schneider, 1998; Turner & Upshur, 2002; Upshur & Turner, 1995, 1999).

Another technique associated with the empirical approach is the use of expert judgment. Raters' scores and their comments provide valuable insights in establishing levels of

performance for performance-based assessment of speaking as well as writing (e.g., Brown, Iwashita, & McNamara, 2005; Cumming, Kantor, & Powers, 2001). The terms *verbal protocol/verbal report* are used to describe the utterances gathered from an individual who is asked to say what he or she is thinking either while completing the task or immediately after completing it (Cohen, 2013; Ericsson & Simon, 1984; Green, 1998). In the case of scale development, a rater is often directed to talk about the features of a response that influenced the score that was given. Many scholars recognize the value of verbal protocols, although problems do include potential reactivity, people's inability to speak fast enough to say everything they are thinking, or their comfort with this type of task (e.g., Asencion, 2004; Green, 1998; Jourdenais, 2001; Thomson & Isaacs, 2008).

In addition to the approach taken in scale development, the method of scoring must also be considered. The two major scoring methods are holistic and analytic. The decision to use one or the other often reflects the use of the test. Holistic scoring is appropriate for contexts in which a large number of test takers are assessed and scoring has to be carried out in a limited amount of time, such as for admission decisions. Holistic scoring has been widely used in a number of high-stakes testing situations around the world (e.g., American Council on the Teaching of Foreign Languages, Canadian Language Benchmarks, Interagency Language Roundtable, and the former TSE). To score holistically is to express an overall impression of the impact of a test taker's performance in one score (Luoma, 2004; McNamara, 2000). In this method, a single score is given to each speech sample based on its overall quality, guided either by the rater's impression or by a rating scale. This single score is aimed at capturing all the features of the speech sample.

There are a few drawbacks, however, regarding holistic scoring. Raters may be influenced by their first impression when they hear an oral sample and then not carefully attend to all criteria specified in the scales (North, 2000). There is also a limitation due to the nature of speaking itself. Speaking performance is complex, simultaneously incorporating the sound system, the repertoire of vocabulary, and the ordering of words together, all to convey meaning (Bygate, 2002; Luoma, 2004). Raters may not be able to balance perceptions of these different language features. For example, a speaker may have a wide range of vocabulary but may mispronounce most words, affecting the intelligibility of his or her response.

For instructional uses, analytic scoring is often preferred as students' strengths and weaknesses can be identified. Analytic scoring requires the rater to provide separate assessments

for each of a number of aspects of performance. The rater, for example, might be required to give separate scores for fluency, pronunciation, vocabulary, and grammar. Use of analytic scales has been advocated in a number of studies in terms of helping raters to balance criteria in the scales and ensuring attention to the same subconstruct (e.g., Brown et al., 2005; Nakatsuhara, 2007; North, 2000; Sawaki, 2007). Illustrating their diagnostic feedback potential, Xi and Mollaun (2006) reported that when the raters assigned analytic scores for each of the three dimensions of the TOEFL speaking rating scale (i.e., delivery, language use, and topic development), rich information was provided regarding strengths and weaknesses of test takers. Another advantage of analytic scoring is that not only can scores be reported for diagnostic purposes but the scores can also be combined and reported, much like holistic scores, so that they can be used for admission or placement decisions (Sawaki, 2007).

Despite these strengths, analytic scoring also presents assessors with some challenges. One challenge involves the extra time and money needed. When coupled with evidence of high correlations between analytic and holistic scores, these practical concerns weigh heavily against analytic scoring. Raters may experience too heavy a cognitive load when trying to score different categories. Also, a halo effect is possible if the raters cannot separate their judgments on one category from the others (North, 2000; Sawaki, 2007).

An approach that reduced analytic scoring's cognitive load was tried out by Upshur and Turner (1995). They developed empirically derived, binary-choice, boundary-definition (EBB) scales. Their scales work like a tree diagram in which the rater answers one yes/no question at a time; his or her answer then leads to the next question, and so on, until a score is reached. In their initial report, the EBB scales were developed by consensus among teachers to score grammatical accuracy and communicative effectiveness of language learners in performing a story-retell task. The scales consisted of a set of explicit yes/no questions relating to students' performance, ordered in levels from overall to specific quality. The rating scales were considered effective because they reflected task demands as well as discourse types and were claimed to be valid because they were derived from authentic samples of performance.

These EBB scales are distinguished from other analytic scales by four major characteristics (Turner & Upshur, 2002; Upshur & Turner, 1995). First, the EBB scales contain fewer descriptors; the rater is supposed to make a choice from two statements that describe particular spoken features, one at a time, until a score is reached for each spoken response. As

reported by Brooks (1957), it is generally easier to discriminate the presence or absence of an ability than it is to discriminate between several levels. It is, therefore, considered simple and practical. Second, instead of describing the midpoint of a scale category as other rating scales do, the EBB scales describe the boundaries between categories; these descriptors are based on the perception of differences rather than similarities. Third, whereas other analytic scales are normally employed in standard test settings, the EBB scales are used in instructional settings. The EBB scales were developed for helping teachers with scoring their students' speaking confidently and reliably. Lastly, the EBB scales allow for diagnosis. When a score is arrived at, teachers can report strengths and weaknesses of a test taker by tracing back to different language features that are included in that particular score.

There are criticisms of the EBB scales. Different aspects of language use are used at different decision points instead of throughout the scales. However, the use of all aspects of language throughout the scales might cause a problem in that the rater, especially a novice rater, may not be able to cognitively focus on all criteria included in descriptors of such scales (North, 2000; Sawaki, 2007; Xi & Mollaun, 2006). The scales were also criticized for their task-specific quality as they were initially developed for a story-retell task. Turner and Upshur argued for the application of the EBB: "The lack of generality of these rating scales is not in dispute, but more general, theory-based rating scales have not been shown to be equally valid for the various task types that empirically derived scales are designed for" (2002, p. 53). It remains to be seen whether EBB-type scales can be used for different tasks. Their empirical basis, ease of use, focus on instructional settings, and potential for yielding diagnostic information make the EBB scales an intriguing possibility for scale development.

Review of TOEFL iBT Scoring Rubric

Turning now to the development of TOEFL iBT's speaking rubric, all three approaches and both methods of scoring are apparent. Beginning with the claim, a TOEFL iBT speaking score is used to infer examinees' ability to produce spoken language in simulated academic tasks reflective of real life in English-speaking universities. The score is derived from raters' holistic consideration of a general description based on three dimensions—delivery, language use, and topic development (ETS, 2008).

In its early development, the speaking construct was defined theoretically as follows: "Communicative competence in oral academic language requires control of a wide range of

phonological and syntactic features, vocabulary, and oral genres and the knowledge of how to use them appropriately" (Butler et al., 2000, p. 2). It was framed by features that account for task difficulty: situational features, discourse features, and test rubric, as well as task and performance conditions (Butler et al., 2000; cf. Iwashita, McNamara, & Elder, 2001). Functionally, the tasks required examinees to produce meaningful speech in communicative situations of varying potential complexity appropriate to academic life such as narrating what happened, sharing an opinion, and comparing and contrasting methods (Rosenfeld, Leung, & Oltman, 2001). Empirically, professionals listened to actual responses and identified salient features of performance at five levels in a series of workshops (Enright et al., 2008). Findings were corroborated in a separate effort, in which raters gave verbal reports on spoken performances: First, raters described their overall impression of a response and second, they mentioned specific performance features that affected their overall impression (Brown et al., 2005). Raters' comments were grouped into four major categories: linguistic resources (grammar, vocabulary, expression, textualization), phonology (pronunciation, intonation, rhythm and stress), fluency (hesitation/pauses/fillers, repetition/repair, speech rate), and content (task fulfillment, ideas, framing). These criteria were similar to those in earlier drafts of TOEFL iBT's rating scales, and indeed they were similar to the three criteria (content, delivery, and language) reported in much earlier research on factors that affect the ratings of speakers (Becker, 1962). One recommendation (that was later adopted) was to provide raters with task-specific information in order to determine content appropriateness. The final form of the integrated speaking scale was developed after a pilot study was conducted having four groups of four professionals trained in English as a second language (ESL) rank responses to six items on a scale of 0–4. Along with a general description, categories in the scale include delivery, language use, and topic development.

Strictly speaking, the TOEFL iBT integrated speaking scale is an example of holistic scoring, as a rater gives one score on a 1–4 scale for a response to a task (0 is given for responses that are too short to score or off topic). However, in order to arrive at that holistic score, raters are instructed to consider the three dimensions of delivery, language use, and topic development (ETS, 2008). A response that is given a score of 4 must be characterized as a 4 along all three dimensions; a score of 3, 2, or 1 must be characterized at that score level in at least two of the three dimensions. So, less strictly speaking, the scale includes aspects of analytic scoring, as it requires that the rater make separate assessments on a number of aspects of performance. For

TOEFL iBT, analyses were conducted to compare holistic with analytic scores. Although shown to be measuring somewhat different constructs, the intercorrelations among holistic score and the analytic scores were generally quite high. It was decided to use the holistic scoring method, taking into consideration the operational concerns of cost and time spent scoring (Enright et al., 2008).

The combination of holistic and analytic scoring could be confusing for novice raters. They may be unsure as to which linguistic features belong to which scoring dimension; this may, in turn, alter the holistic score that is given. Insight into this issue was provided by Xi and Mollaun (2006) as they investigated the utility of analytic scoring for the TOEFL Academic Speaking Test (TAST). Using the same scale and similar tasks as the speaking section of TOEFL iBT, these researchers reported that the raters perceived overlap among the three categories, specifically between language use and delivery, and between language use and topic development. Novice raters may be helped by training that focuses on the distinctions between score points for each of the three dimensions. Xi and Mollaun concluded that holistic scoring is appropriate for operational settings, but that while training holistic raters, examining patterns on the analytic scores may help raters see how performance on the dimensions of delivery, language use, and topic development impact the holistic score. They also encouraged future work with samples of distinct learner groups as a means to providing information that could be considered by the TOEFL program's scoring services. They suggested that in low-stakes situations, such as test practice or for placement and diagnostic purposes, information on the three analytic dimensions could be quite useful.

Dimensions of Speaking

In order to develop analytic rating guides for instruction and for novice raters, the dimensions of delivery, language use, and topic development should be understood both as they are used by TOEFL iBT and as they are studied by applied linguists.

Delivery. In the TOEFL iBT scale, the description of delivery includes the conceptually overlapping qualities of fluency (i.e., fluidity), pronunciation, intonation, and intelligibility. All four of these features have been included in a variety of applied linguistics studies.

Although there are varying perspectives of the concept of *fluency* (Davies et al., 1999; Koponen & Riggenbach, 2000; Kormos & Denes, 2004), Lennon (2000, p. 25) provided a useful definition that works fairly well with the TOEFL iBT scale: Fluency "can be measured both

impressionistically and instrumentally by speech rate, and by such disfluency markers as filled and unfilled pauses, false starts, hesitations, lengthened syllables, retraces, and repetitions." In seminal work with native speakers of English, Goldman-Eisler (1958, 1961a, 1961b, 1961c) provided relevant detail. First, rate is realized by the number of syllables per second. Second, duration is realized by mean length of run. Both rate and duration can be thought of as contributing to the flow of speech, augmenting Lennon's definition of rate. Disfluency is realized by pauses, hesitations, repetitions, choppiness, fillers, and false starts as all refer to disruptions in the flow of speech. Together, the flow of speech and its disruptions combine to give us what we perceive as the speed of talking—fluency. In more recent work with second language learners, fluency is usually judged by the following temporal features of language: (a) speed and density per time unit (or speech rate; Derwing & Munro, 1997; Isaacs, 2008; Kormos & Denes, 2004; Kang, 2010; Munro & Derwing, 2001); (b) breakdown fluency (or number and length of pauses; Derwing, Rossiter, Munro, & Thomson, 2004; Kang, Rubin, & Pickering, 2010; Kormos & Denes, 2004); and (c) repair fluency (or false starts and repetitions; Freed, 2000; Iwashita, Brown, McNamara, & O'Hagan, 2008). In a TOEFL study about the fluency of speakers, raters commented most frequently on speech rate, hesitation, repetition, and repair (Brown et al., 2005).

What is a good speech rate? Increasing speech rates of 1.32, 1.66, 2.02, 2.36, and 2.83 were reported for TOEFL responses at Score Levels 1–5, respectively (Brown et al., 2005). Isaacs (2008) showed that a speaker rated as most comprehensible by native speaker listeners had the fastest articulation rate (mean number of syllables per second excluding pauses; 4.84 syllables per second). In a study by Anderson-Hsieh and Venkatagiri (1994), the mean articulation rate for native speakers was 5.0 syllables per second, the mean rate for the high-proficiency group was 4.4, and the mean rate for the intermediate group was 3.3. However, Munro and Derwing (1995, 2001) cautioned against a faster-is-better interpretation, suggesting a U-shaped function between speech rate and listener's understanding, where the ideal listener rate for English native speakers (NSs) to nonnative speakers (NNSs) is about 4.7 syllables per second; listeners will experience difficulty if the rate of NNSs is too slow or too fast. Overall, speech rate variables have been found to be significant predictors of rated oral performance in nonnative speakers (e.g., Kang, 2010; Kang et al., 2010). Among four speech rate variables, the following two—syllable per second and mean length of run—had a stronger impact on listeners' judgments than others (Hincks, 2005; Kang, 2008; Kormos & Denes, 2004).

Much work regarding disfluencies has involved pauses. In several quantitative studies, pauses have been operationalized as mean length of pauses, number of silent pauses, and total number of filled pauses. Mean length of pause has been found to decrease with increased proficiency—1.0 for low-intermediate NNSs, 0.6 for advanced NNSs, and 0.4 for NSs (Kang, 2008; Kormos & Denes, 2004). Anderson-Hsieh and Venkatagiri (1994) found that the pause rate for intermediate level NNSs was much longer than for either advanced NNSs or NSs. Overall, the number of silent pauses was a strong predictor of the judgments of nonnative speakers' oral performance (Kang, 2008). Other studies examined factors associated with pauses. Goldman-Eisler (1961a, 1961c) found longer pauses for NSs in tasks that were cognitive (monologic descriptions and summaries of meanings) and emotional (psychological interviews) than in tasks that were social (discussions). Fulcher (1993) described eight types of pauses to discriminate between high and low level NNSs. Higher ability students were likely to pause or hesitate when planning how to express a point of view, how to support their views by giving examples, and how to reformulate utterances to clarify their opinions. Lower ability speakers were likely to pause in order to consider grammatical structure or choice of words or to repair their choice.

Pronunciation is the second delivery feature in the TOEFL iBT scale, often collocated with the words *problems* and *difficulties.* Pronunciation can refer to the articulation of individual speech segments but it can also refer to the way individual sounds are perceived by the hearer (Richards, Platt, & Platt, 1992). Derwing and Munro have conducted several studies on intelligibility and comprehensibility of nonnative speakers of English in which they have included discussion of pronunciation, particularly accentedness (Derwing & Munro, 1997, 2005; Munro & Derwing, 1995, 2001). Accentedness is defined as the cumulative auditory effect of pronunciation that identifies a person regionally or socially and includes both stress and intonation (Crystal, 1991).

Intonation is the third delivery feature in the TOEFL iBT speaking rubric. Contrasted from individual segments of sound in language, intonation is classified as *suprasegmental*—sound features that extend over more than one sound segment (Crystal, 1991). Wennerstrom (2000, 2001) used intonation as a general term and one feature of fluency (as do others; see Davies et al., 1999), but in the TOEFL iBT rubric intonation and fluency are separate. In discussions of intonation, two other terms often appear. *Stress* is caused by more air being

pushed out of the lungs, and *pitch* (variations in highness or lowness of sounds) is caused by vibrations of the vocal chords; patterns of pitch in a sentence are what is called *intonation* (Ladefoged, 2006). Although stress and pitch are physiologically different, an increase in air flow from the lungs (stress) also increases the vibration of the vocal chords (pitch) and so stressed syllables often sound higher as well as louder. Wennerstrom associated the higher pitch of stressed syllables with content words containing new information; she said that if, in English, pitch did not vary across words, then discourse cues would be lost and the utterance would sound choppy.

One of the most crucial problems in the intonation patterns of NNSs is an overall narrow pitch range, which makes the identification of prosodic units difficult (Mennen, 1998; Pickering, 2004; Wennerstrom, 1998). Wennerstrom (1994) measured the pitch of NS versus NNS in speaking tasks designed to elicit contrasts between high and low pitch in the areas of information structure and the boundary structure. While the NSs had significant differences in their pitch range to signal these contrasts in all the environments tested, groups of NNSs from Thai, Spanish, and Japanese backgrounds did not show such differences. Speakers from several other first language (L1) backgrounds were also found to have narrower or more compressed pitch ranges than English NS: Chinese (Juffs, 1990; Wennerstrom, 1998), Finnish (Toivanen, Väyrynen, & Seppänen, 2004), Japanese (Taniguchi, 2001), Saudi Arabic (Binghadeer, 2008), and Slovak (Timkova, 2001).

Intelligibility is the fourth delivery feature in the TOEFL iBT rubric. In the past, this term was often used interchangeably with *comprehensibility*. Both terms refer to a listener's understanding of an utterance. Applied linguistics dictionaries for example, define intelligibility broadly as the degree to which a listener can understand a message (Davies et al., 1999; Richards et al., 1992). More recently, Derwing and Munro (1997, 2005) distinguished two types of listener comprehension according to whether a listener actually understands (intelligibility, in which case the listener can accurately transcribe the utterance) or whether the listener perceives difficulty in understanding (comprehensibility, in which the listener judges an utterance on a scale of 1–9 from easy to extremely difficult). It seems the TOEFL iBT rubric adheres to the older definition, as *listener effort* is used along with *intelligibility* in the rubric. In this project, the term *intelligibility* was used in its older, broader sense to mean the degree of listener effort that was

required to understand what the speaker was trying to say; the more intelligible an utterance was, the less listener effort was needed.

Language use. Unlike the mention of four specific features of delivery in the TOEFL iBT rubric, language use is characterized much more globally, including the range (basic versus complex) and control (imprecise versus accurate) of vocabulary and grammar as well as the notion of automaticity.

Addressing vocabulary range in applied linguistics research, *lexical richness* has been thought to indicate effective use of vocabulary in oral speech. The most widely used measure of lexical richness is type-token ratio (TTR). Despite its popularity, this traditional measure has been questioned as not being reliable due to its sensitivity to text length (e.g., Brown et al., 2005; Vermeer, 2000). Vermeer (2000) and Daller, Van Hout, and Treffers-Daller (2003) suggested that a more valid measure of lexical richness in spontaneous speech data would analyze different levels of lexical frequency used by language learners. Brown et al. (2005) included TTR, proportions of low and high frequency vocabulary used, numbers of types, and numbers of tokens in their study of TOEFL speakers. Despite norming the word counts, TTR was found to be higher for responses with lower scores. Difficulty levels of vocabulary had a slight relation to score level. Responses with higher percentages of words from the Academic Word List (Coxhead, 2000) had significantly higher scores, though with a small effect size. Higher scores were also associated with both more words (tokens) and a wider range of words (types).

Addressing vocabulary control, raters in the Brown et al. study (2005) frequently commented on the adequacy of vocabulary, the accuracy/precision of word choice, sophistication/richness/stylistic choices, use of idioms, and reusing input from the prompt. It seems that raters have a better sense of speakers' vocabulary range and control than can be measured objectively.

Grammar-related second language studies often focus on accuracy and complexity in which accuracy reflects the extent to which the speaker can produce the target language in relation to its rules (i.e., control) and complexity reflects the speaker's willingness to use elaborated or difficult language structures (i.e., range).

Measures of grammatical accuracy of learner speech have been operationalized globally (Foster & Skehan, 1996; Skehan & Foster, 1999) and specifically (Iwashita et al., 2008; Ortega, 1999; Wigglesworth, 1997). To determine global accuracy, all errors are considered and

manually coded. Many researchers of spoken language use errors per T-units as a measure of accuracy (a main clause and any subordinate clauses attached to it; Hunt, 1966). Whereas the T-unit was developed for analysis of written language, it has counterparts for spoken language in the C-unit and the AS-unit. All of these are syntactic measures that "allow the analyst to give credit to performers who can embed clauses and hence construct chunks of speech that reflect more sophisticated planning processes" (Foster, Tonkyn, & Wigglesworth, 2000, p. 362). Brown et al. (2005) used the percentage of error-free T-units as their measure of global accuracy; specific errors were counted for tense, third person singular verbs/copula, plural nouns, article use, and prepositions. All measures of grammatical accuracy/error had significant effects on proficiency level of the speakers, although global accuracy had a higher effect size than the specific measures.

In a study of oral proficiency of TOEFL iBT examinees, grammatical complexity was investigated by employing measures adopted from earlier studies of both spoken and written language. Among the measures of T-unit complexity ratio, dependent clause ratio, verb-phrase ratio, and mean length of utterance, only verb-phrase ratio and mean length of utterance showed significant differences across proficiency levels (Iwashita et al., 2008). Another investigation of second language (L2) adult learners' grammatical complexity also included dependent clauses, or subordination, but used a measure adopted from previous spoken discourse research (Norrby & Hakansson, 2007). In contrast to the Iwashita et al. (2008) results, subordination proved a significant indicator of complexity. This finding supports Foster and Skehan's (1996) study on L2 oral performance, illustrating that amount of subordination (i.e., total number of separate clauses divided by the total number of C-units) can be used to measure L2 learners' oral complexity. In Lennon's (1990) longitudinal study of the development of NNSs' oral performance, the use of relative clauses and prepositional phrases per T-unit increased from 18–118% and from 33–51%, respectively.

Three other issues related to grammar and vocabulary are relevant here. Criteria identified as naturally occurring linguistic features in speech are likely to contribute to our understanding of grammatical complexity. Rather than selecting grammatical categories a priori, it may be beneficial to use a more inductive, or data-driven, investigation of salient lexico-grammatical features. As referred to above, many studies of spoken language have included grammatical categories that were found to be important in studies of written language. There is

ample evidence to show that one should be cautious when interpreting such data as when occurrences of grammatical features in spoken language are different from those in written language (e.g., Biber, Johansson, Leech, Conrad, & Finnegan, 1999). Finally, researchers should be open to the possibility that complexity varies according to proficiency level. As pointed out by Rimmer (2006), a data-driven analysis allows the researcher to investigate salient patterns of lower proficiency students who are likely to produce short language at word or phrasal levels as well as patterns of higher proficiency students who are more likely to use various clauses.

Turning to the last feature of language use, *automaticity* relates to the acquisition of a complex skill, such as learning a second language, and its interconnections with attention and effort (Segalowitz, 2003). Learning a new skill requires focused attention, and progress is slow. As we improve, we need to pay less and less attention and performance of the skill becomes easier and easier; performance becomes faster, more efficient, and more stable. Apart from its popular appeal, the claim that second language learning and complex skill acquisition are strongly related lacks a solid research foundation and instrumentation in applied linguistics. Still, the notion has been used in reading in regard to fluent word recognition and also in regard to instructional conditions for grammar structures (Jiang, 2007; Phillips, Segalowitz, O'Brien, & Yamasaki, 2004).

Topic development. In TOEFL iBT's integrated speaking scale, topic development refers to the accuracy and completeness of the content that was provided in the input text(s) and the overall orderly progression of ideas, or coherence, of the spoken text. Because effective language use also facilitates the progression of ideas, TOEFL raters may have experienced confusion as to how to consider cohesive devices (Xi & Mollaun, 2006).

Cohesive devices refer to grammatical and lexical items on the surface of a text that can form semantic and referential connections between the parts of the text (Biber et al., 1999; Halliday & Hasan, 1976). Features such as conjunctions, that can link ideas in a spoken text and thus enhance text coherence and the listener's understanding of the speaker's message, are cohesive devices. Included in Halliday and Hasan's inventory of textual cohesion are five categories of resources, all realized at the lexico-grammatical level—reference, substitution, ellipses, conjunction, and lexical cohesion.

Among the few studies on effects of cohesive devices on language proficiency, two provided support for using referential and temporal markers to distinguish among proficiency

levels of NNS of English. Ejzenberg (2000) focused on the role of cohesive devices (i.e., coordinating conjunctions and adverbial conjunctions) and grammatical devices (i.e., subordinating conjunctions and relative devices) that speakers used to organize their speech. Speakers who received high speaking scores had a higher proportion of cohesive devices to link and organize their monologic talks. In another study Fung and Carter (2007) compared the production of four categories of discourse markers (i.e., interpersonal, referential, structural, and cognitive) by English NS and NNS. Their findings showed that both groups of speakers used discourse makers to organize and structure their speech, but NS used more discourse markers in every category and for a wider variety of pragmatic functions.

Apart from a purely lexico-grammatical viewpoint, discourse structure has been analyzed. Knoch (2007b) made use of topical structure analysis to create a rating scale for coherence in writing. The lowest score level used the following descriptors: "Frequent: unrelated progression, coherence breaks. Infrequent: sequential progression, superstructure, indirect progression" (p. 117). The highest score level had these descriptors: "Writer makes regular use of superstructures, sequential progression; Few incidences of unrelated progression; No coherence breaks" (p. 117). In their analysis of content for TOEFL speaking responses, Brown et al. (2005) counted the number of T-units and clauses and they conducted a qualitative analysis by describing the ideal discourse structure(s) of a task in terms of optional and obligatory moves. They reported that T-units did not uniformly differ across proficiency levels but the number of clauses did gradually increase as proficiency increased. Their qualitative results showed that responses at the lowest score levels used very simple structures—speakers gave their opinion, but without reasons. Responses at the mid level contained more elaborate structures (e.g., more reasons) but they were repetitive and not very clear. Responses at the higher score levels had elaborate structures (reason/opinion/example/reason/example/example/restatement). Similar results were reported by Inoue (2009). In the first part of Brown et al.'s study, raters' most frequent verbal comments for content included task fulfillment, ideas (including amount, relevance, accuracy, and organization), and framing (introduction, conclusion).

The Current Study

Attempting to apply the TOEFL iBT speaking rubric to show progress of learners who are still studying English in school creates a situation similar to that encountered by American foreign language teachers who were using the speaking rating scale developed initially by the

Foreign Service Institute; the lower end of the scale had to be expanded so that students' progress could be shown, resulting in the ACTFL speaking guidelines. This study attempted to expand the TOEFL iBT scale.

When considering TOEFL iBT's use for formative assessment in instructional settings, advantages associated with analytic scoring are particularly appealing as they allow the linguistic components of delivery, language use, and topic development to be broken out by descriptions at different score levels. Phonological, lexical, syntactic, and discourse language features shown to be related to different levels of English proficiency should be described in terms that make sense to teachers, test takers, and raters.

The descriptors of the linguistic features must be grounded in the existing TOEFL iBT rubric so that they will be useful for novice raters, but they must also be presented in more detail to guide instruction. Based on the review thus far, delivery could be explained in terms of speech rate, mean length of run, pause rate, and number of silent pauses for fluency; in terms of choppiness, variation in pitch, and patterns of pitch for intonation; in terms of correct articulation of English segmentals or accent for pronunciation; or in terms of easiness to understand for intelligibility. Language use could be explained in terms of the number of different words, their richness or sophistication, and their correct use for vocabulary, and in terms of a variety of different phrases and clauses, or types of subordination, along with attempts, or fast, efficient performance for grammar. In addition to accurately representing the content, topic development could be explained grammatically by types of conjunctions (e.g., coordinating, adverbial, subordinating) and relative pronouns, or rhetorically by simple or elaborate topical structures or by using introductions to frame a response.

By analyzing language produced by test takers using variables found to be important in other studies as well as looking for naturally occurring patterns in the data and by analyzing the qualities that influence raters' scores, this data-driven approach to scale development could translate key linguistic markers of proficient, academic spoken English into useful, simple terms (Fulcher, Davidson, & Kemp, 2011). Analytic rating guides modeled after the EBB scales' yes/no questions could be simple, reflect underlying language abilities that are needed in academic contexts, and illustrate speaking improvement. In sum, an expanded scale and a series of yes/no questions might provide space needed to clarify linguistic features associated with weak and strong performances. The combination of EBB and data-driven approaches could

provide grounded information to empirically expand TOEFL iBT's speaking rating scales and develop analytic rating guides.

The problem of how to implement this information and expand TOEFL iBT's current holistic speaking rubric for integrated tasks into three effective analytic rating guides for the dimensions of delivery, language use, and topic development was broken down into the following research questions:

> RQ1: Can raters who were extensively trained and experienced using a holistic 4-point scale make meaningful distinctions, when minimally trained, on three 7-point scales (using .5 increments)?
>
> RQ2: For each of the three dimensions, is it possible to identify linguistic features associated with different score levels?
>
> RQ3: For each of the three dimensions, can rating guides be developed that blend the current TOEFL iBT scale with previous research, relevant linguistic features, and criteria that influenced raters' decisions?

Method

This section begins with an explanation of the research design and continues with descriptions of the materials, participants, and procedures. The scope of the study was constrained by several factors including the interests of the TOEFL Committee of Examiners (COE) research agenda, the time available for raters, and cost. That said, the research design and its associated procedures were somewhat complex, so a diagram is presented to help the reader navigate through the phases of the study (see Figure 1).

Research Design

We were motivated by a pragmatic stance in selecting mixed methods to develop our rating guides (Johnson & Onwuegbuzie, 2004; Onwuegbuzie, 2007). The quantitative approach (QUAN) included analyses of linguistic features that were reflected in the TOEFL iBT's integrated speaking rubric or that had been included in previous research. It also included analysis of raters' scores. The qualitative approach (QUAL) used think-aloud protocols of raters' decision-making while they were scoring spoken responses to the integrated tasks. The purpose of using mixed methods was mainly to find convergence between the quantitative and the

qualitative data, providing evidence for the validity of the new rating guides, while also allowing for unique perspectives from one or the other types of analyses (Creswell, 2009; Creswell & Plano Clark, 2011; Greene, Caracelli, & Graham, 1989; Kim, 2013).

Materials

TOEFL iBT includes both independent and integrated speaking tasks. Test takers respond to a prompt by giving their opinions for the independent tasks; for the integrated tasks, they must include specific content in response to reading short texts and/or listening to conversations or lectures that had been developed to resemble authentic materials. Because independent speaking tasks have been the focus of study in previous research on the TSE, the TOEFL research agenda that this study responded to focused on the new integrated tasks.

Four integrated speaking tasks were included in the study. The tasks varied by number and type of input texts and by the rhetorical demand of the prompts. Two of the tasks (3 and 4) included both reading and listening input texts. Task 3 included a short letter in a college newspaper followed by a brief conversation in which a man and woman discussed the opinion of the letter writer; the prompt asked for reasons why the woman disagreed with the reasons expressed in the letter. Task 4 began with a reading on the topic of *groupthink*, followed by an excerpt of a lecture in a business class; the prompt asked for an explanation of the concept of groupthink and its effects, using an example mentioned in a lecture. The other two tasks (5 and 6) included only one listening input text. Task 5 included a conversation between two students, a man and a woman, who were discussing the problem of transporting children on a field trip; the prompt asked test takers to summarize two solutions proposed by the woman and for the test taker's opinion of the preferred solution. Task 6 included a part of a lecture in a psychology class about mathematical capabilities of babies; the prompt asked for a summary of what has been learned about these capabilities, based on the research described. All of the tasks required the speaker to summarize information that was heard, but using somewhat different rhetorical functions: Task 3 required the speaker to state the woman's position, Tasks 4 and 6 required the speaker to identify key features and to relate those to examples that were given, and Task 5 required the speaker to summarize a problem and to support a solution.

Participants

Data set participants. Educational Testing Service (ETS) provided a data set that included audio files for examinees' responses on speaking tasks as well as a data file of their scores on each task; this was a subset of the data from examinees who took the domestically administered TOEFL iBT on November 18–19, 2005. ETS was mainly interested in distinctions between test takers who scored between 2 and 3, as these were the scores the majority of students received and indicated score levels at which finer discrimination might be possible. It was thought that capturing more detailed information about performance at these levels would be especially useful for improving scoring practices and for showing test takers what it takes to move up the scale, from a 3 to a 4, for example.

Two language groups (i.e., Chinese and Spanish) were selected based on their prevalence among TOEFL takers and international students in the United States. This decision was informed by the native language of examinees who took the TOEFL between July 2005 and June 2006 (ETS, 2007a, 2007b) as well as by the number of international students in U.S. colleges and universities in 2006 (Koh Chin & Bhandari, 2006). Data from a total of 68 people whose first language was Chinese (36 examinees) or Spanish (32 examinees) were analyzed (see Table 1). As indicated by the native country code in the ETS data file, Chinese L1 speakers were from five countries, with the majority from China and Taiwan. The Spanish L1 speakers were mainly from Mexico, Central, and South America. The average age of the participants was 26. Most of participants were between 20 and 29, though it can be seen that the Spanish L1 group included somewhat older participants. About 54% of all of the participants were female, with males and females divided similarly across the two L1 groups.

Table 1

Sixty-Eight Participants' Demographic Information

Native language	Native country	Age	Sex
Chinese (36)	China (19), Taiwan (11), Hong Kong (3), United States (2), Canada (1)	20–29 (22) 30–39 (12) 40–49 (2) 50–59 (0)	Male (16) Female (20)
Spanish (32)	Columbia (6), Mexico (6), Venezuela (4), Peru (3), Bolivia (2), Chile (2), Cuba (2), Ecuador (2), Panama (2), Argentina (1), Spain (1), Guatemala (1)	20–29 (14) 30–39 (8) 40–49 (8) 50–59 (2)	Male (15) Female (17)

As detailed in Table 2, there were a total of 231 spoken responses from the integrated speaking tasks (i.e., reading/listening/speaking [RLS] and listening/speaking [LS]). Specifically, for each language group, about 50 responses had scores of 2 (note that there were only 24 Chinese responses available in the data set at this score level). Of these, 24 responses had scores of 2 on the RLS task and another 25 responses from the same test takers have scores of 2 on the LS task. The same scheme worked for scores of 3—there were 50 responses for each language group, with 25 on each of the two types of integrated tasks. In sum, 199 responses were at the 2 or 3 scoring levels.

Table 2

Breakdown of Spoken Responses Used in Study by First Language and Task

Score	Chinese		Spanish		Row total
	RLS	LS	RLS	LS	
1	4	4	4	4	16
2	24	25	25	25	99
3	25	25	25	25	100
4	4	4	4	4	16
Column total	57	58	58	58	231

Note. RLS refers to Reading/Listening/Speaking Tasks 3 and 4; LS refers to Listening and Speaking Tasks 5 and 6.

Although the focus of the study was at Score Levels 2 and 3, where the majority of scores were given, some responses from Levels 1 and 4 were needed to establish the boundaries of the scale and to ensure that expanded scale scores of 1.5 and 3.5 did not overlap with existing scores 1 and 4, respectively. As detailed in Table 2, at Score Levels 1 and 4, 16 spoken responses from two different language groups were selected. For each language group, four responses received scores of 1 on the RLS task and four received scores of 1 on the LS task. Four responses for each language group received scores of 4 on the RLS task and four received scores of 4 on the LS task. Only two speech samples in the data set were at Score Level 0, so this score level was not included in the study (although Score Level 0 would be included in the final scale).

A separate study investigated the roles that task and L1 might play in the development of the rating scales. Based on a subset of 44 participants who completed all four tasks, mixed-model repeated measures ANOVAs were run (adjusting the alpha level for multiple comparisons) using the dimension score and the linguistic features as the dependent variables. For all of these

analyses, there were two independent variables—one repeated (task) and one between (L1). Results indicated that L1 had no effect and task had a small effect on scores or use of linguistic features; these results did not lead us to conclude that the rating guides must be developed separately for each task.

Participating raters. Participants were 35 ETS raters, consisting of 26 females and 9 males. Most of them (43%) were over 50 years old; 28% were between 40 and 49; 26% were between 30 and 39; and 3% were between 20 and 29. Most possessed a master's degree in TESOL, applied linguistics, linguistics, English education, or related field. Of the 35 participants, 34 had experience teaching nonnative speakers of English. All had experience assessing ESL or English as a foreign language (EFL) speaking and the majority were TOEFL-speaking scoring leaders, thus considered to be expert raters. Because these participants were regular raters, only a limited amount of their time was available. In consultation with ETS staff, a total of 5 hours per rater was considered reasonable—2 for training and 3 for scoring and recording their verbal reports (described in detail below).

Procedures

A chart of the study's phases is provided in Figure 1 to help the reader follow the sequence of the procedures and the product(s) that resulted. Each phase is described in turn.

QUAN data collection. In the first phase of the project, ETS was contacted; we were granted permission to use the data set of spoken responses. Sound files of test takers' responses, holistic scores that were given for each response, and demographic information including age, sex, native language, and native country were subsequently sent to us.

QUAN analysis—linguistic and statistical. The first set of analyses was both linguistic and statistical. Spoken responses were coded for linguistic features for each of the three dimensions. Both sound files and transcripts were used to code these features.

For delivery, two of the four features found in the TOEFL iBT rubrics (i.e., fluency, intonation, pronunciation, and intelligibility) were analyzed quantitatively. Considering available resources including computer programs for phonological analysis, fluency, and intonation were included in this phase; we decided that the features of pronunciation and intelligibility would be adequately evaluated by our human raters as part of the qualitative analysis. As Kang (2008, 2010) has summarized, whereas in the past much of the work on the intelligibility of speech

focused on pronunciation of individual sounds, more recently the focus has shifted to the role that speech rate, pauses, and intonation play.

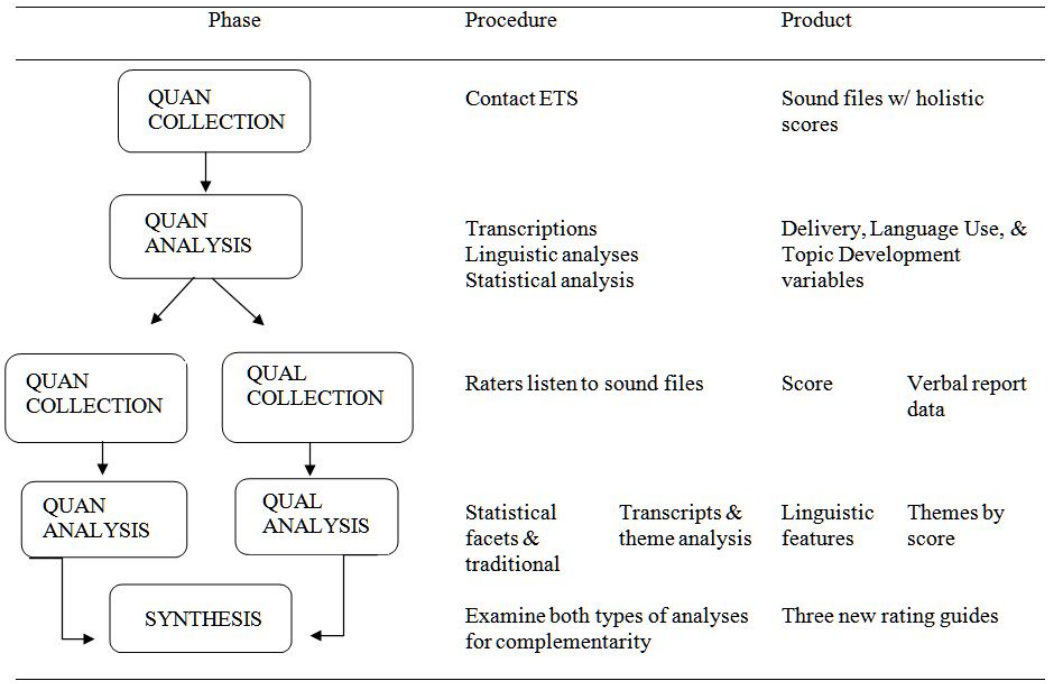


Figure 1. **Major phases, procedures, and products of the mixed methods research design.**

Fluency was operationalized by three variables—syllables per second, mean length of run, and silent pauses; intonation was operationalized by pitch (i.e., overall pitch range). These variables belong to accented measures in pronunciation called *acoustic fluency*, the best predictor of rated oral performance (Kang, 2008, 2010; Kang et al., 2010) and reflect the descriptors listed in the TOEFL iBT speaking rubrics. The computer program Praat (Boersma & Weenink, 2009) was used to measure these four phonological variables for each of the 231 speech samples:

- Syllables per second were operationalized as the total number of syllables produced in a given speech sample divided by the amount of total time required to produce the speech sample (including pause time).
- Mean length of run was calculated as the average number of syllables produced in utterances between pauses of 0.1 seconds and above.
- Silent pauses were calculated as the average number of silent pauses per minute.
- Overall pitch range was calculated from points of fundamental frequency (F0) maximum and minimum in the 1-minute utterances where the prominent

segments were identified by F0 peaks, which distinguished each prominent syllable from the surrounding content. As the project included male as well as female participants, pitch was transformed into semitone (Couper-Kuhlen, 1996).

For language use, the TOEFL iBT rubric referred to speakers' range and control of vocabulary and grammar, as well as automaticity. In this quantitative phase, automaticity was not included due to our inability to operationalize it adequately. Each of the 231 audio files was transcribed. This spoken corpus consisted of 25,667 total words, with 12,129 words in the RLS tasks and 13,538 words in the LS tasks; this corpus was the basis of the quantitative analysis for both language use and topic development.

Based on resources including computer-assisted analysis programs available and previous research (e.g., Brown et al., 2005; Vermeer, 2000), vocabulary was operationalized in terms of lexical richness and vocabulary range:

- Vocabulary richness was calculated as a proportion of low and high frequency vocabulary used in [the transcripts of] each spoken response using the web program *VocabProfile* (Cobb, 2002).
- Vocabulary range was measured by a TTR using the computer concordancing program, *MonoConc* (Barlow, 2000).

Because of the lack of specificity of particular structures in the TOEFL iBT rubric, two approaches were used for operationalizing grammar. Based on the literature, the first approach included measures of grammatical accuracy and grammatical complexity:

- Grammatical accuracy was measured by summing the number of error-free C-units.
- Grammatical complexity was examined by counting occurrences of complement clauses, adverbial clauses, relative clauses, and prepositional phrases:
  - Occurrences of complement clauses were derived from a combination of occurrences of *that* complement clauses and *to* complement clauses.
  - Adverbial clauses included three types of subordination conjunctions: causative (e.g., *because*), conditional (e.g., *if*, *unless*), and others (e.g., *except*, *until*).
  - Relative clauses represented both *wh* relative clauses and *that* relative clauses;

- Prepositional phrases were examined by looking at phrases beginning with prepositions followed by noun phrases.

Coding was done both manually and automatically. Grammatical accuracy was computed manually. Two researchers marked printouts of the transcripts for clauses (i.e., C-units) and then counted the number of clauses that did not contain grammatical errors; coding agreement for grammatical accuracy was .84. Grammatical complexity was measured automatically. The transcripts were automatically tagged for grammatical (lexical and morphosyntactic) features. The tags were rechecked by the researchers using the reference grammar *Longman Grammar of Spoken and Written English* (Biber et al., 1999). It should be noted that this coding system has been used to describe features of both written and spoken language, for both L1 and L2 English writers and speakers.

The second approach to operationalizing grammar was data-driven as it was not theoretically motivated but rather empirically motivated; that is, data-driven grammatical complexity was examined by counting occurrences of the lexico-grammatical features generated by Biber's (2001) tagging program. This approach resulted in a number of grammatical variables that emerged from about 70 lexico-grammatical categories tagged.

For topic development, the TOEFL iBT rubric referred to the accuracy and completeness of the content provided in the speech and its overall progression of ideas. The framework of the analysis was based on Halliday and Hasan's (1976) lexico-grammatical model of cohesion with some adjustments for clearer operational definitions of the conjunction and collocation relations (Biber et al., 1999; Tanskanen, 2004; see Appendix A).

The variables of topic development were reference devices for cohesion, conjunction devices for cohesion, and inclusions of introduction and key ideas. Though not explicitly referred to in the TOEFL iBT scale, presence of an introduction had been found to be related to the TOEFL holistic score in Brown et al. (2005) and so was included.

- Reference devices were operationalized by summing occurrences of personal pronouns, demonstrative pronouns, and comparative adjectives.
- Conjunctive devices were operationalized by summing occurrences of addition, apposition, result, contrast, transition, and summation.
- Inclusion of introduction was operationalized as whether or not there was an introduction in a response (a dichotomous variable).

- Inclusion of key ideas was operationalized as a number of key ideas from the scoring key that were present in each text.

Inclusions of introduction and key ideas were counted by two coders. Initial agreement for presence of introduction was .75; coders discussed each response and came to 100% agreement. Using a scale of 0–3, coders' correlation for the number of key ideas was .65; exact agreement was 51% and agreement within one idea was 91%. Again, coders came to agreement on the number of key ideas.

Apart from the three variables coded manually (grammatical accuracy, inclusion of introduction, and number of key ideas), the counts of these grammatical features used to operationalize language use and topic development were normalized to 100 words because that was about the average length of responses. The counts of the tagged features were imported to SPSS 17.0 and initially screened to see if all grammatical features occurred in the corpus. Descriptive statistics (mean and standard deviation) were calculated to find nonoccurring features. Features without counts were deleted from the list. Redundant features whose counts overlapped with other features were excluded from the study to avoid fraudulent counts. For example, some of the categories were sums of other separate categories. Descriptive statistics were computed for all of these variables. Then, for each set of variables, correlations were run with the holistic score provided by ETS. This initial set of analyses served two purposes during this phase of the study. First, it provided an objective measure of actual linguistic features in responses that had been predicted in the literature or had been included in the TOEFL iBT rubric. Second, it provided us with talking points for the verbal reports; in other words, the verbal reports were semistructured based on our empirically established variables.

QUAN + QUAL data collection. As shown in Figure 1, in the next major phase of the study both quantitative and qualitative data were collected. Raters were recruited via emails informing them briefly about the project. After a training session, raters listened to one set of spoken response scoring and providing verbal reports as to their reasons for giving a certain score. Then, raters listened to a second set of responses, only scoring. This phase generated both scores (QUAN data) and verbal reports (QUAL data) and is explained in more detail in the following paragraphs.

All participants in the study read an informed consent form and confidentiality request and replied to the email before taking part in the project. They also informed us their choices of

training dates and times they would be able to participate in a training session. After that, a rater packet was mailed to individual participant. The packet included training date and time, headphone/microphone (for those who needed a set), and materials used in the training session and the actual scoring session for the project (i.e., Elluminate Website for Research Study Training, TOEFL iBT integrated speaking rating scale, four TOEFL iBT integrated tasks including reading passages and/or listening transcripts, prompts, key points, and topic notes, and a CD with listening files for each task).

The raters received 2 hours of online training via video-conferencing with Elluminate, Inc. (2008) on distinguishing among speaking performances and giving verbal reports. The first part of the training was devoted to scoring, specifically, how to distinguish among seven levels of performance (not only holistically, but also just for delivery, language use, and topic development). The researchers familiarized the raters with the idea of an expanded TOEFL iBT integrated speaking rating scale beginning with an overall impression (i.e., holistic). When expanding a scale, a question arises as to the optimal number of steps, or points, in a scale. Too few do not allow for an indication of student progress, but too many may result in underutilized scores, representing apparent categories that cannot be reliably discriminated. Dating back to the work of Miller (1956), the number seven has been often found to represent a reasonable, if not optimal, number of steps in a scale, including speaking scales (e.g., Becker & Cronkite, 1965; Hofmans, Theuns, & Mairesse, 2007; Preston & Colman, 2000). Raters were asked to listen to a set of responses representing seven scores (1, 1.5, 2, 2.5, 3, 3.5, and 4) from an RLS task. They were asked to rank order these responses, based on their knowledge of the TOEFL iBT integrated speaking scale. After that, they compared their rankings with those of the other raters in their training session and discussed salient features of responses at adjacent levels. After working on holistic scores, the raters were asked to rank other sets of seven responses in turn—for either delivery, language use, and/or topic development.

Each set of sample responses had been selected from the entire TOEFL data set by the researchers. First, the data set was examined and eight responses were randomly chosen from an RLS task, two of which represented scores of 1, 2, 3, and 4 each (assigned from ETS raters using the TOEFL iBT integrated speaking rubric). After that the researchers individually listened to and rank ordered the responses using the TOEFL iBT integrated speaking rubric. Then they compared their rankings and discussed salient features of the responses at the different levels to

reach consensus. When their rankings were discrepant, more responses from the data set were examined and the same process of finding consensus of sample responses representing the seven score levels was repeated. This was done to ensure that the examples of speech at each score level were clearly distinct from one another. Selection of the seven example responses was carried out for overall impression (holistic scoring), followed by delivery, language use, and topic development; there were, thus, 28 sample responses in total.

The second part of the training session focused on the raters' producing verbal protocols. This part began with an introduction, the objective of the verbal protocol technique, and guidelines for how to produce verbal protocols. The researchers answered any questions regarding the production of protocols. Next, the raters were asked to practice producing verbal protocols via a practice website. The raters clicked on a link in order to practice with the website. They were asked to play and listen to the spoken response, score it (1, 1.5, 2, 2.5, 3, 3.5, 4), and record their verbal protocols regarding anything that influenced their scoring decisions. They were also instructed to listen to their recording to make sure that they had successfully recorded their protocols. Any concerns regarding the scoring, protocol, or technological issues were discussed or solved before the raters scored spoken responses and produced and recorded verbal protocols in the actual study. Both parts of the training session were recorded for development of benchmark responses and descriptions. About 10 training sessions were held between November 2008 and January 2009; the number of raters per session varied between one and five.

About 2 weeks after the training sessions, each rater received a unique link to a website (developed by staff at Northern Arizona University for this project) that contained the following pages: scoring instructions; benchmark responses (based on those speech samples that had been agreed on during the training) and descriptions for scoring; and, as shown in Figure 2, controls for listening to the sound files, for scoring on each of seven expanded score levels (i.e., 1, 1.5, 2, 2.5, 3, 3.5, and 4), and for recording their verbal responses.

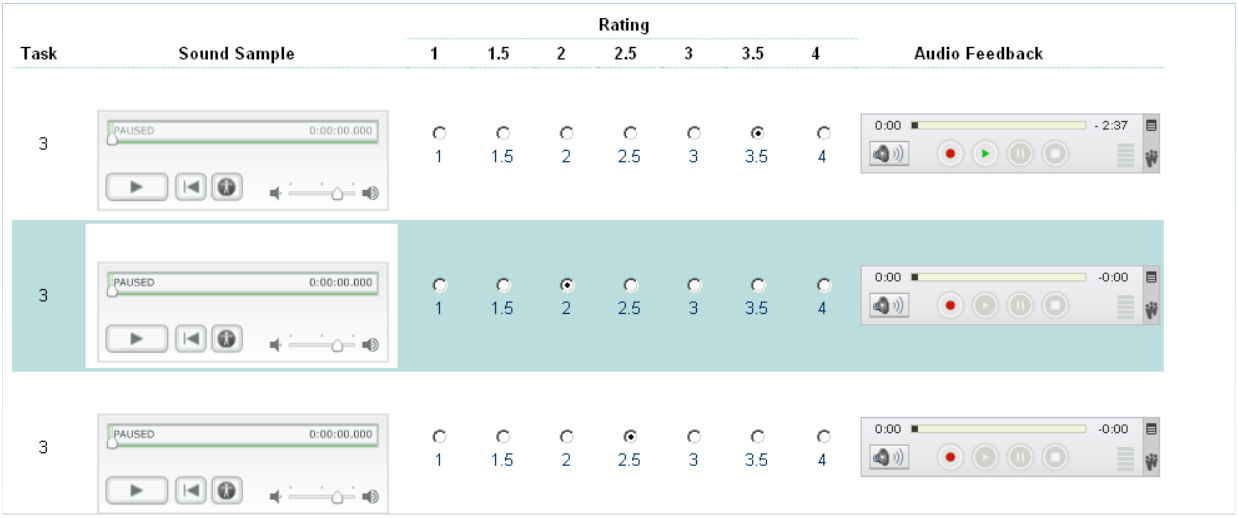


Figure 2. **Rater scoring page scrolled down to show sample, rating, and recording.**

A total of 231 responses were double scored for overall quality and each dimension; thus a total of 1,848 scorings were conducted (231 responses x 2 scorings x 4 dimensions). Each of the 35 raters evaluated about 53 spoken responses subdivided into 13 or 14 speech samples for overall quality and each of the three dimensions. These were systematically arranged for double scoring across levels, tasks, and first language group, taking into account data linkages that would be necessary for subsequent analyses. The benchmark responses and descriptions were selected based on raters' consensus scoring in the training sessions for each of seven expanded score levels (i.e., 1, 1.5, 2, 2.5, 3, 3.5, and 4) for overall quality, delivery, language use, and topic development. Features in the TOEFL iBT rubrics, previous analysis of salient linguistic features, relevant literaure, and raters' comments informed the development of a brief list of talking points for the verbal protocols for each dimension, designed to encourage raters to describe their reasons for their scoring decisions. For example, raters were encouraged to reflect on their scoring decisions in regard to hesitations, pauses, pronunciation, intonation when scoring for delivery, to talk about range and accuracy of vocabulary and grammar for language use, and to address coherence, task relevance, framing, and accuracy of ideas for topic development. Raters were encouraged to include any and all reasons for their scoring decisions. The raters were expected to spend 4 minutes for each response (i.e., 1 minute for listening to a response, 1.5 minutes for scoring, and 1.5 minutes for providing a think-aloud protocol).

Scoring was divided into two parts. For each part, raters were given half of their speech samples (26 or 27 total). In the first part, raters both scored and completed think-aloud protocols

whereas in the second part they only scored. In the first part, raters listened to the first set of responses (6–7) and score each holistically; they recorded their verbal reports. This was included so that raters would begin with a familiar task; the holistic scores and verbal reports are not included in this report. The next set of responses (6–7 different responses) was scored for delivery, each score followed by a verbal report. Then, a third set of responses (a different 6–7) was scored for language use; again, each score was followed by a verbal report. The final set of responses (another 6–7) were scored for topic development and followed by verbal reports.

In the second part, the raters listened to and scored 26 or 27 new speech samples, again subdivided into four sets of 6 to 7 responses (holistic scoring, delivery, language use, and topic development). There were no verbal reports in this part. Raters were expected to spend 2.5 minutes on each response (i.e., 1 minute for listening to a response, 1.5 minutes for scoring). Each rater spent about 3 hours completing the scoring and verbal protocol tasks.

QUAN and QUAL analysis. As shown in Figure 1, the next phase of the study included both QUAN and QUAL analyses, which are explained below in separate sections.

For QUAN analysis**,** two raters had scored each speech sample; their scores were examined separately in some analyses and were averaged in others.

To address Research Question 1 (whether minimally trained raters could make meaningful distinctions when using a 7-point scale rather than a 4-point scale), raters' scores were examined separately. Frequency tables allowed us to examine the degree to which the raters actually used the expanded scale. A Facets analysis (version 3.64.0) allowed us to look at the consistency of the raters, their severity, the difference between tasks, and the meaningfulness of the expanded levels of the scale (Myford & Wolfe, 2003, 2004). Traditional estimates of inter-rater reliability were also computed. Because only the first rater gave his/her reasons in verbal reports (due to time constraints), the influence of verbal reports on ratings was checked.

To address Research Question 2 (which linguistic features were associated with different score levels), correlations and multiple regressions were computed to examine the inter-relationships among the dimension scores and the linguistic features; for these QUAN analyses, raters' scores were averaged and analyses were conducted separately for each of the four tasks because so many of the participants completed more than one task. Due to the relatively small number of participants for each task (56–58), the results of the multiple regression analyses were considered exploratory. Descriptive statistics were also computed for the significant features at

each of the score levels in the cases when both raters agreed on the score. For each dimension, about 20% of discrepant responses were listened to by a third rater (one of the researchers) and the most frequent score was used. Examination of these means and standard deviations would potentially allow us to characterize a score level and to distinguish between score levels.

For the QUAL analysis addressing Research Question 2, the verbal report data were analyzed at each score level. This analysis involved the steps of transcription, coding, developing propositions, and developing themes. The researchers believed that this qualitative approach was more suitable for capturing the thoughts of the raters than quantizing the frequency of their similar ideas, although both types of analysis are accepted for convergent designs (Creswell & Plano Clark, 2011). The recorded verbal reports were transcribed using consistent transcription conventions (Biber et al., 2004; Edwards & Lampert, 1993). The transcripts were verified against the original data by the researchers before being coded. Coding of the verbal reports was done by the authors and was based on pieces of information in the transcripts. This coding system followed Miles and Huberman's (1994) framework, involving underlining pieces of information (words, phrases, or sentences) that related to any of the three dimensions of the scale (i.e., delivery, language use, topic development). Three labels were given to the underlined segment: the dimension, the characteristic feature of that dimension, and a score. Immediately after a coder underlined and labeled the pieces of information, she also wrote down as many ideas as occurred to her about the labels and their relationships (i.e., memoing). Once this process was completed for one spoken text, the coder looked over all of the underlined segments, labels, and memos and came up with a summary statement, or a *proposition*, for each of the characteristic features. Then, each coder examined all of the verbal reports for a given score level and summarized the propositions; these were called *themes*. Finally, the two coders compared and discussed their themes, grouped related themes, and agreed on a final set of themes at each score level for each dimension.

During our pilot study, characteristic features of each dimension had been identified through coding the language of the raters that seemed to refer to one of the three dimensions found in the TOEFL iBT rubrics. The coding agreement between two raters was estimated at .83 (Poonpon, 2009). For delivery, the coding scheme included the categories of general, fluency, hesitation, clarity of speech, pauses, pace, pronunciation, intonation, and listener effort required. For language use, it included a general category, vocabulary range or control, vocabulary

accuracy or errors, grammar range or control, and grammar accuracy or errors. For topic development, it included general, topic development with a clear connection to the prompt or clear connections of ideas, organization of ideas, repetition of ideas, overall rater understanding, completeness of ideas, accuracy of ideas, and vague or unclear ideas.

During this study, each dimension also had an *other* category that expanded the coding scheme. When a coder thought that the characteristic features needed to be expanded, both coders met, examined the data, and agreed either to add more categories or to code the questionable piece of information with an existing category. For delivery, other categories that were added were sustained, intelligibility, natural, reference to prompt, confidence, repetition, self-correction, sentence boundary, and automaticity. For language use, length, communication of ideas, repetition, intelligibility, attempt, cohesive devices, automaticity, self-correction, and confidence were added. For topic development, prompt comprehensibility, attempt/effort, length, and detail were added. Intercoder reliability for each dimension was .84 for delivery, .76 for language use, and .73 for topic development.

Synthesis. The final phase of the procedures depicted in Figure 1 was to examine the quantitative and the qualitative results in order to address Research Question 3: Can rating guides be developed for each of the three dimensions? In this phase, the 7-point scale that had been used in all previous analyses was changed to an 8-point scale because the score of 0 was reintroduced. (Recall that there were only two speech samples with a score of 0 in the ETS data set, and so a decision had been made to exclude the score of 0.)

Having independently analyzed the quantitative and qualitative data, a procedure was needed to decide what information to compare across the two data sets. Establishing a set of decision rules that drew on previous work by (Turner & Upshur, 2002; Upshur & Turner, 1995, 1999), findings were examined to determine points at which speakers' responses could be divided into two successively smaller groups based on a series of yes/no questions.

First, we looked for features that separated low scores from high scores or the lower half of scores (0–2) from the upper half of scores (2.5–4). The statistically significant variables were considered in relation to the themes, and both were compared with findings from previous research studies and theories. We termed this the Level 1 decision point and it resulted in two groups. Next, we looked for features that could be used for yes/no decisions for each of these two groups. The Level 2 decision points subdivided the lower half (scores 0–1 versus 1.5–2) and

the upper half (scores 2.5–3 versus 3.5–4), and resulted in four groups. Finally, we looked for features that could be used for yes/no decisions to separate each of the four groups into score levels (0 versus 1, 1.5 versus 2, 2.5 versus 3, and 3.5 versus 4); these were the Level 3 decision points.

To foster a positive effect when the rating scale was used, we wrote these decision points as yes/no questions, in which a yes answer would be associated with desirable speaking features, which in turn would lead to a higher score. Quantitative results, qualitative results, previous research, and theoretical claims were subjectively compared to find evidence of complementarity, but also to be open to insights shown only by one of the approaches.

Results

The results are presented in response to each of the three research questions, which are restated for the convenience of the reader.

RQ1: Can raters who were extensively trained and experienced using a holistic 4-point scale make meaningful distinctions, when minimally trained, on three 7-point scales (using .5 increments)?

One of the main purposes of the study was to develop expanded scales by providing space for in-between scores. Speech samples that had already received scores of 1, 2, 3, or 4 using the TOEFL iBT holistic scale were rescored on 7-point scales (1, 1.5, 2, 2.5, 3, 3.5, 4) separately for delivery, language use, and topic development. Two raters scored each speech sample. In order to answer this research question, we begin with the descriptive statistics of the scores, followed by a frequency distribution to see if all score points were being used. Then, the results of the Facets analyses allowed us to examine the functioning of the 7-point scales and the consistency of the raters. Traditional interrater reliabilities are also provided. Finally, as raters gave verbal reports for the first half of their scores but not the second, the effect of verbal reports is reported.

In order to get a better sense of the score data, descriptive statistics and histograms for the original TOEFL iBT score and the three dimension scores given by raters in the study were compared (Table 3). Recall that the selected data set for this study (TOEFL iBT scores) included 16 responses at Score Levels 1 and 4 in order to establish a baseline and ceiling and that the majority of the speech samples were at Score Levels 2 (n = 99) and 3 (n = 100) so that

distinctions would be better able to be made at the 2–3 bands where the majority of test takers score.

Table 3

Descriptive Statistics for Scores Used in the Study

Score	*N*	M	SD
Delivery	231	2.63	.66
Language use	231	2.56	.68
Topic development	231	2.55	.76
TOEFL iBT	231	2.51	.73

The descriptive statistics were similar across all scores. The means for the three dimension scores in the study ranged between 2.55–2.63, with that of delivery being a bit higher than the others. The standard deviations were between .66–.76, with topic development scores showing somewhat more variability than the others.

Each of the 231 speech samples was scored on each dimension by two raters. One issue that needed to be addressed was whether the raters actually used all of the expanded scale points. As shown in Table 4, raters made use of all of the .5 score expansions for each dimension. About 50% of the scores given were at 1.5, 2.5, and 3.5.

Although all score points were used, the question remained as to whether the scores on each of the scales were properly ordered (Myford, 2006, pp. 83–90). Two statistics from the Facets analysis address this issue of the proper functioning of the scales. *Average measures* shows the average proficiency measures at a score level; these should increase as the scores increase, showing that speakers with higher scores have more of the quality we are measuring. *Most probable from* shows the lowest proficiency measures that a score should be observed at; these should also increase as the scores increase. *Low* indicates the most probable score at the lower end of the scale; if the value is *never* or *no,* some scores are underused, indicating a problem with the scale. As shown in Table 5, for all three scales, the values of both average measures and most probable from increased as scores increased, indicating that the scores on the scales were properly ordered.

Table 4

Frequency of Scores for Each Dimension

Score	Delivery		Language use		Topic development	
	n	%	*N*	%	*N*	%
1	10	2	15	3	31	7
1.5	47	10	52	11	50	11
2	83	18	98	21	87	19
2.5	112	24	108	23	95	21
3	118	26	98	21	95	21
3.5	64	14	68	15	69	15
4	28	6	23	5	26	6
Total	462	100	462	100	462	100

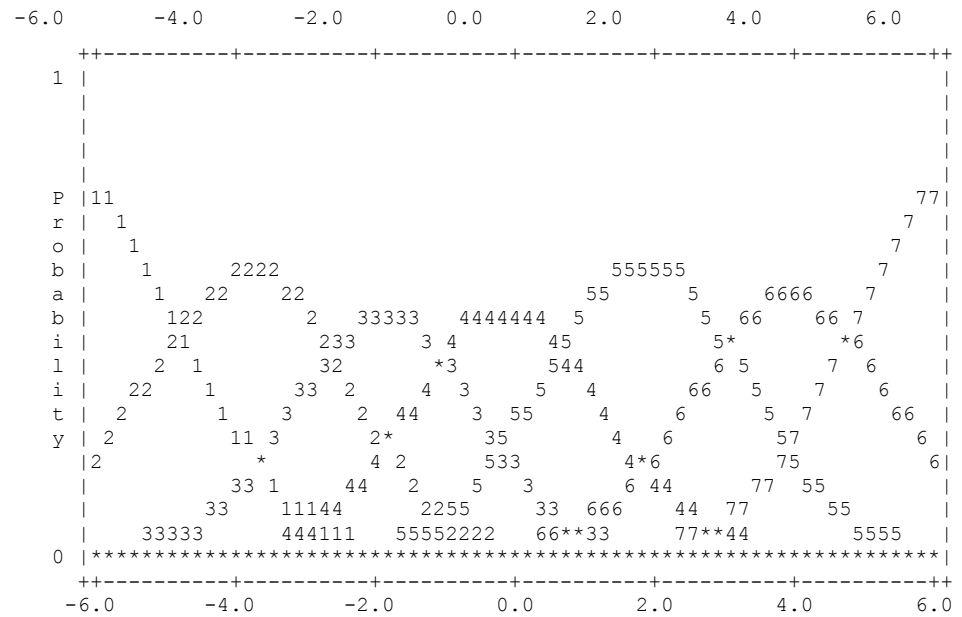
Table 5

Facets Results on Functioning of Dimension Scales

Score	Delivery		Language use		Topic development	
	Average measures	Most probable from	Average measures	Most probable from	Average measures	Most probable from
1	-3.91	low	-4.02	low	-3.42	low
1.5	-2.68	-4.77	-3.01	-4.76	-2.02	-3.10
2	-1.47	-2.61	-1.77	-3.01	-.94	-1.92
2.5	.00	-1.05	-.64	-1.30	.04	-.58
3	1.43	.62	1.51	.43	.88	.39
3.5	3.44	3.04	3.82	3.04	2.10	1.76
4	4.47	4.77	5.10	5.59	2.65	3.59

Another question about the ordering of the scales that can be asked is, Were the score levels clearly distinguishable? This question was answered by examining the graphic output of the category probability curves for each of the three scores. Myford (2006, p. 86) stated that "the chief concern is whether there is a separate peak for each category or not. If some categories never become most probable (and thus to not have separate peaks), then that may suggest that

categories should be collapsed." As shown in Figure 3, for the delivery scale there were separate peaks for each score on the 7-point scale, indicating that the score levels were clearly distinguishable; due to space constraints, this is the only such figure presented, but similar patterns were found for language use and topic development scores.



Note. Score levels 1, 1.5, 2, 2.5, 3, 3.5, 4 represented by numbers 1-7; x axis = proficiency scale in logits

Figure 3. **Category probability curves for delivery.**

Results of Facets analyses addressed two other questions. First, were the raters different in the severity with which they scored the speakers in the study? Two statistics were used to answer this question, the rater separation index, which indicates the number of distinct levels of rater severity (which we hope is near 0 to indicate that raters are interchangeable) and fixed chi-square, which tests whether the raters were equally severe after adjusting for measurement error (which we hope is not significant if the raters are interchangeable). The meaning, use, and interpretation of these and other Facets statistics were guided by Myford (2006).

As we can see from the rater separation index in Table 6, there were about two levels of rater severity for the delivery score, whereas there were closer to three levels for language use and topic development. (Note that reliability here refers to the reliability of the rater separation.) The significant chi-square results support the finding that raters were different in severity on all

dimension scores. So, the answer to the first question is, Yes, raters did differ in severity, although not extremely.

Table 6

Facets Statistics for Rater Severity

Scale	Rater separation index	Reliability	Fixed chi-square	*df*	Significance
Delivery	2.21	.67	100.3	34	.00
Language use	3.24	.83	181.1	34	.00
Topic development	2.62	.75	134.2	34	.00

Second, did each rater use the scale consistently? The answer to this was found by examining the output regarding the mean square infit. For low stakes applications such as we find with this study, Myford (2006) suggested using bounds of .50 to indicate overfit, or less variation than expected, or overuse of a small range of scores, and 2.00 to indicate misfit, or more variation than expected (McNamara, 1996). Myford wrote that misfit is more serious than overfit, as it indicates an unreliable rater.

Considering that 35 raters scored on three different dimensions with limited rater training, there were relatively few cases of misfit and overfit. Using the mean square infit values greater than 2.00, there was one misfitting rater for delivery (R7) and using the mean square value of less than .50, there were four overfitting raters (R2, R6, R10, R35). For language use, there were three misfitting raters (R4, R5, R13) and four overfitting raters (R8, R17, R21, R24). For topic development, there was one misfitting rater (R26) and two overfitting raters (R27, R32). These results indicate that more than 80% of the raters showed the expected amount of variation when scoring on all three dimensions, with about 3% to 9% showing more variation than expected and about 6% to 11% showing less variation than expected. Because none of the misfitting or overfitting raters were the same on the different dimensions, all raters were retained in subsequent analyses.

Next, traditional interrater reliability statistics indicated adequate agreement, considering that raters were introduced to the practice of giving seven scores rather than four during the training. As seen in Table 7 correlations between two raters ranged between .68 and .72 across

the three dimension scores. Using the 7-point scale, there was exact agreement between the raters about 35% of the time, but over 80% agreement within 1 point.

Table 7

Interrater Reliability Statistics for 7-Point Scale

Scale	Pearson correlation	Percentage of exact agreement	Percentage of agreement within 1 point
Delivery	.68	37%	81%
Language use	.72	36%	84%
Topic development	.71	34%	80%

Because the first rater gave verbal reports but the second rater did not, a question arose about what effect the verbal reports might have on the ratings. As shown in Table 8, the mean scores by both raters were similar except for language use in which the raters giving the verbal reports (Rater 1) were slightly more severe than the raters who only scored (Rater 2) by one tenth of a point. Overall, it appears that the verbal reports had little, if any, effect on the leniency/severity of the ratings.

Table 8

Paired T-Tests Between Scores for Raters Giving Verbal Reports or Not

Scale	Rater	Verbal report	M	SD	*t*	*df*	*p*
Delivery	1	+	2.63	.74	-.46	230	.65
	2	-	2.64	.70			
Language use	1	+	2.51	.74	-2.68	230	.01
	2	-	2.61	.74			
Topic	1	+	2.53	.82	-1.14	230	.25
development	2	-	2.58	.83			

In sum, the data indicated that raters were able to use the 7-point scale to make meaningful distinctions among speaking samples with adequate consistency.

RQ2: **For each of the three dimensions, is it possible to identify linguistic features associated with different score levels?**

This question was addressed using both quantitative and qualitative analyses, which are presented in turn, for each dimension. Beginning with the quantitative results, linguistic features for each dimension were described statistically (see Appendix B), correlated with each other and

the dimension score assigned by raters, regressed on the dimension score to estimate their relative importance, and finally summarized at each score level for each task.

For the qualitative results, theme analyses were conducted for the verbal reports on delivery, language use, and topic development. The themes reflected the salient spoken features that raters attended to while scoring at a level, using categories previously established and adding other categories as needed, based on comments that could not otherwise be coded. For each of the seven score levels, final themes are reported according to the score given by the rater who provided the verbal report. As shown in Table 9, the number of different raters was greatest in the midrange of scores, because that is where the majority of the sample was selected (see Table 2). The final themes agreed on by both researchers were used to describe features at one level and to distinguish scores at adjacent levels. Recall that each of the 35 raters provided verbal responses on six to seven spoken responses for each dimension. Recall also that each verbal response was coded according to categories of features, so that any one verbal response could have several comments on several different categories. These were summarized into one or more propositions. Each proposition was then compared across those of other raters at the same score level and was ultimately included as a final theme.

Table 9

Number of Different Raters Who Commented on Each Dimension at Each Score Level

Scale	1	1.5	2	2.5	3	3.5	4
Delivery	6	15	19	17	19	13	11
Language use	4	14	17	17	13	16	8
Topic development	9	14	17	16	17	14	11

Typically, themes were represented by at least five comments, except at Score Levels 1 and 4, where there were as few as two comments. For example, in delivery at Score Range 3, the first theme of sustained, fluent, and generally clear speech had 31 comments, whereas at Score Range 1 the theme of unclear speech had five comments. If themes were identified by a large number of comments at one level of performance, we included related themes at adjacent levels even if they reflected fewer comments to mark similarities or differences in perceived abilities.

Due to space constraints, only one rater's comment is listed for each theme at a given score level; it is hoped that the reader can "hear" the raters' voices. (The full set of raters' comments is available upon request.) Understand that not all responses at any level exhibited exactly the same features. Recall that the raters were by themselves, thinking aloud into a microphone; as a result, their quoted comments in the following sections contain a number of disfluencies and language mistakes, such as would be found in the speech of anyone talking to him- or herself.

Delivery: Quantitative results. Beginning with the quantitative results, correlation analyses revealed that there were significant, positive correlations between the delivery scores given by the raters and the counts of syllables per second, mean length of run, and overall pitch range derived from the linguistic analysis. On all four tasks, statistically significant positive correlations were found between delivery score and syllables per second (.66, .73, .59, and .69, $p < .05$). The correlations between delivery score and mean length of run across tasks were .60, .76, .78, .74 ($p < .05$). For overall pitch range, the correlations with delivery score were .60, .67, .54, .57 ($p < .05$). For all but Task 3, there were significant, negative correlations between delivery scores and number of silent pauses, ranging from -.12 (*ns*), -.27, -.35, -.30 ($p < .05$). These results indicate a positive relationship between rate of speech and intonation with scores but a negative relationship between disfluencies (i.e., silent pauses) and scores. There were no correlations among the four acoustic variables greater than .75, with most ranging between .40 and .69 indicating that multicollinearity was not an issue. It appeared that that while these variables were related, they were measuring somewhat different subconstructs of delivery.

In order to examine the contributions of each acoustic variable to the delivery score in more detail, a multiple regression analysis was conducted for each task. These exploratory results in Table 10 show that mean length of run, syllables per second, and overall pitch range (and silent pauses for Task 3) accounted for .59 to .70 of the variance of the delivery scores.

Table 10

Multiple Regression of Measures on Delivery Score by Task

Task	*N*	Measure	B	SE B	ß	R^2 change	R^2
3	56	Syllables per second	.76	.09	.82	.44	
		Silent pauses	-.02	.01	-.42	.15	.59

Task	*N*	Measure	B	SE B	ß	R^2 change	R^2
4	58	Mean length of run	.22	.06	.40	.58	
		Syllables per second	.31	.11	.31	.08	
		Overall pitch range	.07	.03	.25	.04	.70
5	56	Mean length of run	.40	.04	.78	.61	.61
6	56	Mean length of run	.32	.08	.51	.55	
		Syllables per second	.34	.14	.31	.04	.59

Note. Excluded variables: Task 3 = mean length of run, pitch; Task 4 = silent pauses; Task 5 = syllables per second, pitch, silent pauses; Task 6 = pitch, silent pauses.

Looking for patterns, descriptive statistics were computed for each acoustic variable at each of the proposed scores (1, 1.5, 2, 2.5, 3, 3.5, and 4) using only the responses on which two raters agreed (see Table C1 in Appendix C). Although no clear distinctions could be made between each score level for each variable, some trends were evident. In what follows, the ranges of the mean values are given to provide the reader with a sense of these trends. Syllables per second was the one variable that increased at every score level. Average rates at Score Level 2 and below ranged from 2.07 to 2.67, a relatively slow rate, characteristic of beginner-level speakers. Rates at Score Level 2.5 ranged from 2.75 to 3.00, characteristic of low-intermediate speakers. Rates at Score Levels 3 and above, ranging from 2.72 to 3.91, were relatively faster; though still not considered advanced, these rates are characteristic of intermediate to perhaps high-intermediate ESL speakers.

Mean length of run also increased. At Score Levels 2 and below mean length of run ranged from 2.25 to 2.93 and can be considered in the beginner range. Rates at Score Level 2.5 ranged from 3.05 to 3.41, characteristic of low-intermediate speakers. Rates at Score Level 3 ranged from 3.98 to 4.70, considered intermediate. Those at Score Level 3.5 ranged from 4.49 to 5.50, considered intermediate to high-intermediate.

Pitch range increased somewhat between Score Levels 2 (ranging from 3.91 to 5.08) and 2.5 (4.70 to 6.51), again between 2.5 and 3 (ranging from 6.56 to 7.79), and most noticeably between levels 3 and 3.5 (6.62 to 10.20).

Finally, one could argue for a slight reduction of silent pauses between score points 2.5 (ranging from 54.00 to 59.60) and 3 (ranging from 39.20 to 51.22), but the range between Score Levels 1.5 to 2.5 did not vary much, nor did the range between Score Levels 3 to 4.

Delivery: Qualitative results. For each of the seven delivery score levels, final themes are reported. The majority of the verbal comments were classified according to hesitations, pauses, pace, pronunciation, intonation, and whether the response was fluid, clear, hard/easy to listen to, intelligible, and sustained. Some other categories addressed naturalness, confidence, and self-correction.

Score Level 1. Three final themes from the raters' verbal protocols described delivery at the lowest Score Level, 1. Overall, though, raters made very few comments. First, it was very difficult for many raters to understand anything other than occasional words. For example, Rater 35 (R35) commented, "Really can't say that I would understand much ah… other than er… words here and there throughout the response." Second, the speech was unclear. Commenting on different responses at this score level, R27 said, "unclear words mumbled individual words clear." Third, the responses were very short. As R8 said, "There was just not enough uh speech sample to be judged."

Score Level 1.5. Four themes characterized scores at Level 1.5. Raters thought that these responses had lots of hesitations, false starts, and long pauses and were very choppy. Secondly, many raters thought that responses at this score level were not clear; raters commented on mumbling and lack of enunciation. The third theme related to poor pronunciation and lack of intonation that made the response difficult to understand and required a lot of listener effort. Finally, raters thought that the responses contained some language and that they were sustained in places, but not consistently so. Some raters indicated that this was the reason the response was given a score of 1.5 rather than a score of 1. Some of the raters' comments follow, listed in order of the four themes:

R34: at the word level, there are lots of pauses

R22: a few mumbled words that don't really make much sense

R15: his pronunciation is is awful he eats his words again

R21: she does attempt to umm… sus… provide a sustained response

Score Level 2. Five themes emerged from the analysis of the verbal reports at Score Level 2. First, raters thought that some chunks of speech were fluent and sustained, but not throughout the response. Second, similar to a score of 1.5, quite a bit of listener effort was still required. Third, the pacing of responses was inconsistent, resulting in lots of choppiness. Fourth, intonation was inconsistent and mostly monotone. Finally, although some responses at Score Level 2 were generally clear with adequate pronunciation, more often raters commented that speakers' pronunciation and accents affected their ability to understand. Examples of raters' comments follow, again listed in order of the themes:

R34: some phrases which are, which are spoken fluently but ah… it's not consistent

R18: the significant listener effort comes in though when we try to piece together those words (unclear words) to interpret the meaning

R28: she starts out much slower than she finishes. She picks up some steam and she umm… is able to ah… er… be a little more successful in the response towards the end

R8: his variation and pitch is not well he's monotone in his speech so intonation is poor

R11: she has a consistent accent which basically affects every word

Score Level 2.5. Six themes emerged from the raters' verbal reports. First, pacing was inconsistent, with lots of choppiness, as it was at Score Level 2. Second, although the speech was fluent in parts, it was not fluent throughout. Third, the clarity of the response was affected by some hesitancy and false starts. Fourth, pronunciation was generally clear but with a noticeable accent. Fifth, raters thought that there was some use of intonation although speech was sometimes monotone. Sixth, raters commented that overall intelligibility was fair, even though some words or phrases were hard to understand and required some, but not a lot of, listener effort. Here are some of the raters' thoughts about 2.5 delivery responses:

R32: about the ha- half way into the response, there… er… is noticeable problems with pacing

R10: there are quite a few lapses in terms of the fluidity

R12: some hesitancies that interfere with listening to this response

R8: he has some word level pronunciation issues

R10: the intonation fit what she was talking about uh but that to was was inconsistent to the degree that I did not give this a three

R18: she's fairly intelligible for the most part

Score Level 3. Five themes emerged for responses scored 3. First, raters thought that the speech was sustained and generally fluent and clear. This is different from what they thought about many responses at Score Level 2.5. Second, raters thought that there were still some issues with pacing, but not a lot of false starts, hesitations, fillers, and choppiness, with pauses—but pauses in natural places, as if the speaker were recalling information. Third, pronunciation still caused some problems for raters, particularly due to accent. Fourth, raters thought that good intonation patterns were used to emphasize meaning and contrast, but these were inconsistent, sometimes awkward, and sometimes missing. Fifth, although raters thought that responses required some listener effort, they also thought that, in general, they were able to comprehend the speaker's message:

R11: the speech is basically sustained

R17: seemed not to be hesitations of thinking but more of hesitations within his speech

R3: there definitely was some pronunciation problems that were affected by his uh first language

R5: I did find the speaker to be quite ah adept at uh varying his speech pattern to to provide emphasis throughout his utterances

R2: there was not really any difficulty in understanding what she was trying to say

Score Level 3.5. Six themes characterized raters' scoring considerations at this second-highest level. First, raters thought that responses were fluent, clear, and easy to understand. Some also mentioned that speech was generally sustained, though in comparison to lower score levels, this was rarely commented on. Second, raters commented on only minor pronunciation problems. Third, raters noticed some pauses and hesitations and commented that perhaps the speakers were searching for words. Fourth, raters thought that the pacing of responses was occasionally a bit choppy. Fifth, on a more positive note, raters commented that

intonation was generally good, although it was infrequently mentioned. Sixth, the last theme for Score Level 3.5 was that these responses did not require much listener effort:

R23: she almost had native fluency

R4: his um pronunciation of individual words was near native-like

R26: the reason it wasn't a 4 is because she umm… spent some time searching for words. Some of the pauses preceded ah… it seemed like there were like voc-, it was vocabulary retrieval pauses

R23: she had some umm…pacing issues that made her er…a little bit of chop to her delivery

R18: he's got a very natural and native-like (pronunciation and) intonation

R34: it's easy to listen to

Score Level 4. In general, raters made few comments, though a bit more than at Score Level 1. There were five themes associated with this score level. First, raters thought that the speech was very fluent, clear, and sustained. Second, raters did comment on some hesitancy and pauses, stating that perhaps the speaker was collecting his/her thoughts. Third, raters thought that pacing and intonation were natural and native-like. Fourth, raters thought the responses had good pronunciation with little accent. Fifth, and finally, raters commented that responses were vey intelligible and that no listener effort was needed:

R24: very fluent speaker

R15: they were natural pauses that uh even uh a a native speaker might potentially make

R4: his pacing was very natural as was the intonation

R33: her intonation particularly was I thought umm native like

R3: I felt she was very native-like in her (pacing and) pronunciation

R15: few pauses that she had in there did not uh interfere with her intelligibility

The delivery themes that emerged from the analysis of the comments that the raters made as their verbal reports are summarized in Table 11.

Table 11

Summary of Final Themes for Delivery

Score	Theme
1	• very difficult to understand anything other than occasional word • unclear speech • very short response
1.5	• lots of hesitations, false starts, long pauses, choppy • not clear • monotone, poor pronunciation make speech hard to understand and required lots of listener effort • somewhat sustained in places but not consistent
2	• some chunks fluent and sustained but not throughout • listener effort is required • pace is inconsistent; lots of choppiness • intonation inconsistent and mostly monotone • accented pronunciation often affects listener's ability to understand
2.5	• pacing inconsistent—lots of choppiness • speech is fluent in parts but usually not throughout • some hesitancy with some false starts affect clarity • generally clear pronunciation, though often associated with accent • some use of intonation but inconsistent; sometimes monotone • some words and phrases hard to understand and require listener effort; overall intelligibility fair
3	• speech is sustained, generally fluent, and clear • some, but not a lot of false starts, hesitations, fillers, choppiness, and pauses, with pauses in natural places—perhaps recalling information • accented pronunciation caused some problems • good intonation patterns used to emphasize meaning, contrast but inconsistent—sometimes awkward, monotone • requires some listener effort, but generally nothing much interferes with listener's comprehension
3.5	• fluent, clear, easy to understand, speech was sustained though rarely commented on • only minor pronunciation difficulties • some pauses/hesitations perhaps looking for words

Score	Theme
	• occasionally a bit choppy • intonation generally good • didn't require much listener effort
4	• very fluent, clear, and sustained • some hesitancy, pauses perhaps to collect thoughts • good pronunciation; little accent • very intelligible; no listener effort needed • pacing and intonation were natural, native-like

Language use: quantitative results. In total, 12 variables were significantly correlated with language use scores ($p < .05$). Two variables, error-free C-units and word count were significantly correlated with language use score all four tasks. One variable was significantly related to language use score on two tasks—passives. Nine other variables were related to the score on one task: TTR, primary verbs, lexical verbs (all negatively related), prepositional phrases, adjectives, relative clauses, *wh*-words, adverbs of degree, and adverbs of place.

Looking specifically at the correlations by task, the variables significantly related to language use scores for Task 3 were error-free C-units (.63), passives (.28), word count (.62), and adverbs of degree (.39). The variables for Task 4 were error-free C-units (.44), prepositional phrases (.44), word count (.55), and adjectives (.33). For Task 5, the variables were TTR (-.35), error-free C-units (.56), relative clauses (.35), word count (.63), *wh*-words (.28), and lexical verbs (-.29). For Task 6, the variables were error-free C-units (.58), passives (.40), adverbs of place (.37), word count (.54), and primary verbs (-.29). None of these variables were intercorrelated above .60 on any task.

In exploratory regression analyses, these variables accounted for .52–.62 of the variance of the language use scores across tasks as shown in Table 12. Error-free C-units (a measure of linguistic accuracy) and word count were significant predictors of the language use score on all four tasks.

Table 12

Multiple Regression of Measures on Language Use Score by Task

Task	*N*	Measure	B	SE B	ß	R^2 change	R^2
3	56	Error-free C-units	1.37	.24	.51	.40	
		Word count	.01	.00	.49	.22	.62
4	58	Word count	.01	.00	.39	.30	
		Prepositional phrases	.13	.03	.37	.12	
		Adjectives	.09	.03	.31	.11	
		Error-free C-units	1.03	.31	.30	.08	.61
5	57	Word count	.01	.00	.55	.38	
		Error-free C-units	1.41	.25	.49	.23	.61
6	58	Error-free C-units	1.27	.38	.36	.34	
		Word count	.01	.00	.35	.09	
		Passives	.23	.07	.31	.09	.52

Note. Excluded variables: Task 3 = passives, adverbs of degree; Task 5 = TTR, relative clauses, lexical verbs, *wh*-words; Task 6 = adverbs of place, primary verbs.

Descriptive statistics of the language use features were computed at each of the proposed scores (1, 1.5, 2, 2.5, 3, 3.5, and 4). Error-free C-units seemed to increase between score points 2, 2.5, and 3. Word count increased at each score level. Relative clauses and passives were used more at Score Level 3.5 than at any other level. Use of prepositional phrases seemed to increase between Score Levels 2.5 and 3. Apart from these features, patterns across score levels were not clear for TTR, adverbs of place or degree, or lexical verbs (see Table C2).

Language use: qualitative results. Most of the verbal comments that raters made while scoring were categorized according to range and accuracy of vocabulary and grammar or of language use in general. Occasionally, raters commented on the length of the response, communication or intelligibility, repetition, cohesive devices, automaticity, self-correction, and self-confidence. Themes at each score level are enumerated and followed by example rater comments, listed in the same order as the themes. Again, recall that raters were being recorded as they were thinking aloud, and so their quoted comments contain some ungrammatical language.

Score Level 1. Three themes emerged from analysis of the comments at Score Level 1. There was very little to score, and what language there was showed limited range and many errors, as shown in the rater comments:

R18: very little of anything here to evaluate

R1: there are gross errors all over the place

R29: he just didn't seem to have very much vocabulary at all

Score Level 1.5. Five themes characterized the verbal comments at this score level. First, in general, the language was considered poor. Second, most raters commented that in general there was a limited range of both vocabulary and grammar, but better than a response at Score Level 1. Third, word choices were quite basic. Fourth, there was use of phrases and simple sentences. Fifth, there were many errors at a basic level in morphology, tense, and gender:

R1: the language really was was awful

R15: I bumped him up from a one to a one point five because he puts a lot of language in there

R31: her word choice and vocabulary is, is really limiting her in this response

R7: it does have a lot of complete sentences but they're very very simple and short sentences

R34: grammatical problems in tense and gender…these weaknesses definitely keep her from communicating much on the topic

Score Level 2. Three themes summarized the comments at Score Level 2. On the negative side, basic vocabulary was often used incorrectly. Second, simple sentences were mainly used but with many errors, such as subject–verb agreement errors. Third, on the positive side, chunks of automaticity were evident at the phrasal level:

R7: it's very limited vocabulary a lot of uh uh the words are quite basic

R1: this candidate stays in the simplest I mean the simple you know the simplest of structures no complex structures are really even attempted

R28: errors like subject–verb agreement is throughout

R11: she has a bit of automaticity of expression at simple phrasal level but not throughout

Score Level 2.5. The comments at this score level were grouped into five themes. First, raters thought that speakers did use basic vocabulary, but in addition attempts were made to use more advanced or complicated vocabulary. However, raters also commented that vocabulary sometimes was vague, inaccurate, or awkward. Third, raters thought that speakers had good control of simple grammatical structures and that they made attempts to use more complex structures. However, raters also commented on a fairly wide range of grammatical errors including articles, prepositions, and tense. Fifth, raters noticed that some speakers displayed a little automaticity at the phrasal level and with simple sentences, but they thought that at this score level it was lacking overall:

R35: this speaker is choosing some more advanced vocabulary words

R31: in the latter half of the response, vague word choices

R11: a little difficult for her she uses a few complex sentences but mostly there's simple sentences

R35: a lack of being able to use conjunctions to join together er… phrases

R2: it was um not very automatic or it was automatic only at a fairly simple level

Score Level 3. Four themes emerged at Score Level 3, characterized with somewhat more complexity and fewer errors than Score Level 2.5. First, raters thought that vocabulary use went beyond the basic level, though it was still considered limited. Second, raters mentioned speakers' good control of simple structures and some use of complex structures. Third, raters did comment on some grammar errors. Fourth, raters thought that speakers had good automaticity in many parts of their responses:

R34: the vocabulary is ok ah…it is simple, but er…you know… ah… not that simple either

R1: it used simple structures well…attempts some complex structures

R5: there's too many errors for a 3.5 or a 4

R29: some good automaticity with his grammatical structures

Score Level 3.5. Three themes characterized comments about scores at this level. First, a wider range of vocabulary was used by speakers than that commented on at Score Level 3, with errors mainly associated with use of more sophisticated or complex words. Second, raters thought that speakers used simple structures and many complex structures with only minor errors. Third, a few raters mentioned the speakers' automaticity:

R14: maybe a little bit of simplicity but ya know he he also exhibited uh more sophisticated vocabulary as well

R11: she has a problem when she tried to used a sophisticated word

R31: this person uses umm… basic s- sentence structures very well. … she uses more complex structures fairly well.

R23: her language use seemed very automatic

Score Level 4. Four themes characterized raters' comments at this score level. First, they mentioned that speakers had a good range of vocabulary with precise, effective, and correct word choice. Second, raters thought that speakers had good control of both simple and complex structures. Third, automatic use of grammar and vocabulary was mentioned by a few raters. Finally, as was noted earlier at Score Level 4 for delivery, raters commented on natural or native-like use, now for vocabulary:

R11: he has sophisticated vocabulary and range in word choice

R7: she includes uh passive verb forms

R11: very good control of basic and complex grammatical structures

R11: automatic expression of relevant ideas

R22: a lot of word vocabularies just very much on target with what a native speaker would say

The themes for each score level for language use are summarized in Table 13.

Table 13

Summary of Final Themes for Language Use

Score	Theme
1	• little to score • limited range of language • many errors
1.5	• generally, language use was poor • limited range of vocabulary and grammar • basic word choices • some phrases, simple sentences • many grammatical errors at basic level such as morphology, tense, gender
2	• basic vocabulary often used incorrectly or repeated from the prompt • used simple sentences with many errors such as subject–verb agreement • chunks of automaticity evident at phrasal level
2.5	• used basic vocabulary as well as attempted more advanced/complicated • vocabulary sometimes vague, inaccurate, awkward • good control of simple sentences with a few attempts at more complex structures • some grammatical errors with articles, prepositions, tense • displays a little automaticity at simple/phrasal level but overall it is lacking
3	• vocabulary beyond the basic level but still limited with some errors • good control of simple structures and some use of complex structures • occasional grammatical errors such as third person, verb tense, articles, prepositions • good automaticity found in many places
3.5	• minor errors with basic vocabulary and some advanced, sophisticated uses with both success and awkwardness at times • many complex structures and simple structures with only minor errors • automatic use of vocabulary and grammar
4	• good range of vocabulary with precise, effective, correct word choice • good control of a range of both simple and complex structures • automatic use of grammar and vocabulary, but mentioned less than in 3.5 • natural use of vocabulary; comments not found in 3.5

Topic development: quantitative results. Comparisons between the topic development score and the four variables (key ideas, conjunction devices, reference devices, and introduction)

for each of the four tasks showed that the number of key ideas and the presence of an introduction had statistically significant correlations on all tasks ($p < .05$). For Task 3, the correlation with key ideas was .67; other significant correlations ($p < .05$) were with conjunction devices, .41, and the presence of an introduction, .35. Significant correlations with score for Task 4 were with key ideas (.71) and introduction (.32). Significant correlations for Task 5 were with key ideas (.63), reference devices (-.30), and introduction (.43). For Task 6, there were significant correlations with key ideas (.74), reference devices (-.26), and introduction (.74).

Exploratory multiple regressions (Table 14) showed that this set of independent variables accounted for between .39–.59 of the variance of the topic development scores. Although other variables were related to scores on Tasks 3 and 6, the number of key ideas was clearly the most important variable in this set. This finding was supported when examining the descriptive statistics that were computed at each score level (see Table C3).

Table 14

Multiple Regression of Measures on Topic Development Score by Task

Task	*N*	Measure	B	SE B	ß	R^2 change	R^2
3	56	Key ideas	.56	.08	.64	.46	
		Conjunction devices	.19	.05	.35	.12	.58
4	59	Key ideas	.72	.09	.71	.51	.51
5	58	Key ideas	.68	.11	.63	.39	.39
6	58	Key ideas	.47	.08	.61	.55	
		Introduction	.37	.17	.24	.04	.59

Note. Introduction added as dummy variable. Excluded variables: Task 3 = introduction; Task 4 = introduction; Task 5 = introduction, reference devices; Task 6 = reference devices.

Topic development: qualitative results. The majority of the raters' comments were classified in response to seven questions: (a) was the response connected to the prompt/task; (b) were the ideas connected; (c) was the response developed; (d) was the response coherent; (e) were the ideas complete; (f) were the ideas accurate; and (g) did the speaker repeat him/herself. Other comments were classified according to whether the speaker understood the prompt and

rater's knowledge of the prompt. As with the other two dimensions, a summary of themes at each score level is followed by illustrative comments.

Score Level 1. The comments of the raters at this score level were captured by three themes. First, raters said that they had little relevant information on which to assign a score. Second, raters thought that the speakers had no key points or details, but did have some inaccurate information. Third, raters thought that the prompt had not been understood or that the speaker repeated the prompt:

R18: very little here to evaluate

R19: it was inaccurate with regards to the information

R10: I have no get no indication from this woman's response that she understood the conversation between the two students about the afterschool program

Score Level 1.5. Raters' comments at this score level reflected four themes. First, speakers were able to respond to the prompt, but only at a very superficial level. Second, raters commented on the inadequacy of the ideas in reference to the key points and details. Third, a few raters mentioned that there was little framing or that it was ineffective. Fourth, raters did not think that the ideas were well connected:

R12: she really doesn't talk much to the topic

R23: the two points that she, he makes are inaccurate

R15: he jumps right in with the example there's no context given

R29: he has no transitions, so he just sort of spits out the ideas

Score Level 2. Four themes reflected raters' comments. First, many said that in responding to the prompt, the speakers did include some relevant information, but that ideas were vague, including some inaccurate information. Second, raters mentioned that details were few. Third, raters thought that speakers generally had only occasional connections of ideas that made the response hard to understand. Fourth, overall raters mentioned that there was little or no framing of the response:

R20: the cost of the sculpture but that was even vague, even though I understood, it was still vague

R 16: it did convey some relevant information…however she didn't um really go into many of the reasons

R32: I would say slightly less than fair perhaps inadequate amount of cohesion ah…throughout the response

R23: the framing was er…of course not there

Score Level 2.5. Four themes captured raters' comments at this score level. First, raters thought that speakers responded to the prompt by including most key ideas, but like responses at Score Level 2, these ideas were still stated at a general level. Second, raters thought that the amount of detail given was fair, but not well developed. Third, raters thought that some ideas were well connected, but that others were not. Fourth, a few raters mentioned that the speaker had an introduction:

R35: the speaker does hit the key points but in a very general way

R29: she doesn't really give any details about the example. So umm… she gives some general ideas but nothing very specific or any details.

R35: this doesn't have …I think really coherent, a fairly coherent expression to get it into that 3 range.

R18: the speaker has a very nice introduction, which is the summary of the results of the experiment

Score Level 3. Four themes characterized raters' comments at this score level. First, raters thought that most responses included relevant content that included most if not all key points. Second, raters thought that the amount of detail speakers gave was adequate, but could have been much more developed. Third, raters said that the response was basically coherent, but that speakers had problems using transitions between ideas; the progression of ideas was not always well connected. Fourth, raters said that some speakers did try to frame their responses by using an introduction:

R12: this person covers the topic pretty well uh by explaining the situation and and uh also gives his own opinion

R1: I would say adequate detail's provided

R2: he went right into that without any transition without any connection uh at all from the previous idea

R15: he starts out with a with a good introduction

Score Level 3.5 There were four themes at this score level. First, raters thought that responses generally fulfilled most, if not all, task requirements. Second, raters said that speakers used a good amount of detail and examples that still could have been more developed. The third theme reflected clear and coherent progression of ideas. Fourth, raters thought that some speakers did a good job of framing of the response, but that others did not:

R4: the number of correct ideas I guess was fine

R5: it's lacking in in sufficient detail to award a four

R2: there was a pretty clear progression of ideas

R33: the guy did a goo- ver- job of framing in it

R33: taking too long framing the problem

Score Level 4. Four themes summarize raters' comments at the highest score level. First, raters unequivocally thought that the topic was addressed and key points were included. Second, raters all commented positively on speakers' use of a lot of details and examples. Third, raters said that responses were coherent with a clear progression of ideas. Fourth, raters commented on good framing:

R7: it covers not only all the content of the uh of the uh material but also adds her own comments and uh opinions about the situation being discussed

R24: he volunteered a lot of extra information that you don't usually hear…he went into great detail

R35: very effectively and coherently explain the process of groupthink

R25: it was clearly framed

The themes for topic development across score levels are summarized in Table 15.

Table 15

Summary of Final Themes for Topic Development

Score	Theme
1	• Little relevant information • No key points or details; inaccuracies • Prompt not understood or repeated
1.5	• Able to answer prompt at superficial level • Ideas not clear about details, key points • Generally poor framing if any • Ideas not well connected
2	• Response to prompt includes vague reference to some key points; some inaccuracies • Details not well developed • Occasional, though poor connection of ideas and irrelevant information make the response hard to understand • Little or no framing of response
2.5	• Speaker responds to prompt and includes most key ideas but at a general, sometimes vague level • Not many details • Most ideas not well connected • Limited use of introduction
3	• Mostly relevant content including key points • Amount of detail adequate but could be better • Basically coherent response though progression of ideas not always well connected; problems with transition between ideas • Some framing of response attempted in introduction
3.5	• Generally provides key points required by the prompt • Good amount of detail and examples, could be better • Clear and coherent progression of ideas • Framing of response attempted with varying degrees of success
4	• All key points required by prompt addressed • Very good amount and quality of details and examples • Clear, coherent progression of ideas • Good framing

RQ3: For each of the three dimensions, can rating guides be developed that blend the current TOEFL iBT scale with previous research, relevant linguistic features, and criteria that influenced raters' decisions?

To answer this research question, the current TOEFL iBT scale was used as a starting point. We then synthesized the quantitative and qualitative results with previous findings to develop the new rating guides. The guide developed for each dimension is explained in turn.

Delivery. All of the quantitative pronunciation measures and many of the raters' comments in the study referred to aspects of fluency. Responses given higher scores were associated with higher number of syllables per second, good pace, longer mean lengths of run, and sustained speech. Responses given lower scores were associated with less of these qualities and more disfluencies. For pronunciation, high scores were associated with clear, well articulated pronunciation with little accent and only a few errors, whereas low scores were associated with accented pronunciation that contained errors and was sometimes mumbled, not clearly articulated. For intonation, higher scores were associated with use of varied pitch, stress, and intonation. For intelligibility, high scores were associated with no or very little listener effort to understand what the speaker was trying to say, whereas low scores required listener effort and the speaker's meaning was often difficult to understand. These results are similar to those reported by Anderson-Hsieh and Venkatagiri (1994), Brown et al. (2005), Derwing and Munro (1997), Kang (2008, 2010), Kang et al. (2010), Kormos and Denes (2004), and Munro and Derwing (2001). Figure 4 illustrates the flow of the yes/no paths at the Level 1, Level 2, and Level 3 decision points for the delivery scores.

Reviewing all of the results, three features were considered for the Level 1 decision point: speech rate (the number of syllables per second and mean length of run), use of some intonation (overall pitch range), and the amount of listener effort. This was a crucial decision point—the point at which speakers would be divided into the upper half of the scale or the lower half. Three reasons favored using the amount of listener effort (interpreted here as degree of intelligibility or ease of understanding) as the feature that would divide responses into upper and lower halves. First, the degree of intelligibility clearly distinguishes scores of 2 from scores of 3 on the TOEFL iBT speaking rubric for delivery. Second, there was a clear difference in the themes about listener effort derived from the verbal protocols between scores of 2 and below and those at 2.5 and above. Third, this choice underscores the importance of communication; we

speak to convey our ideas, opinions, or feelings to others—the listener should be able to concentrate on the substance of the speech rather than its formulation.

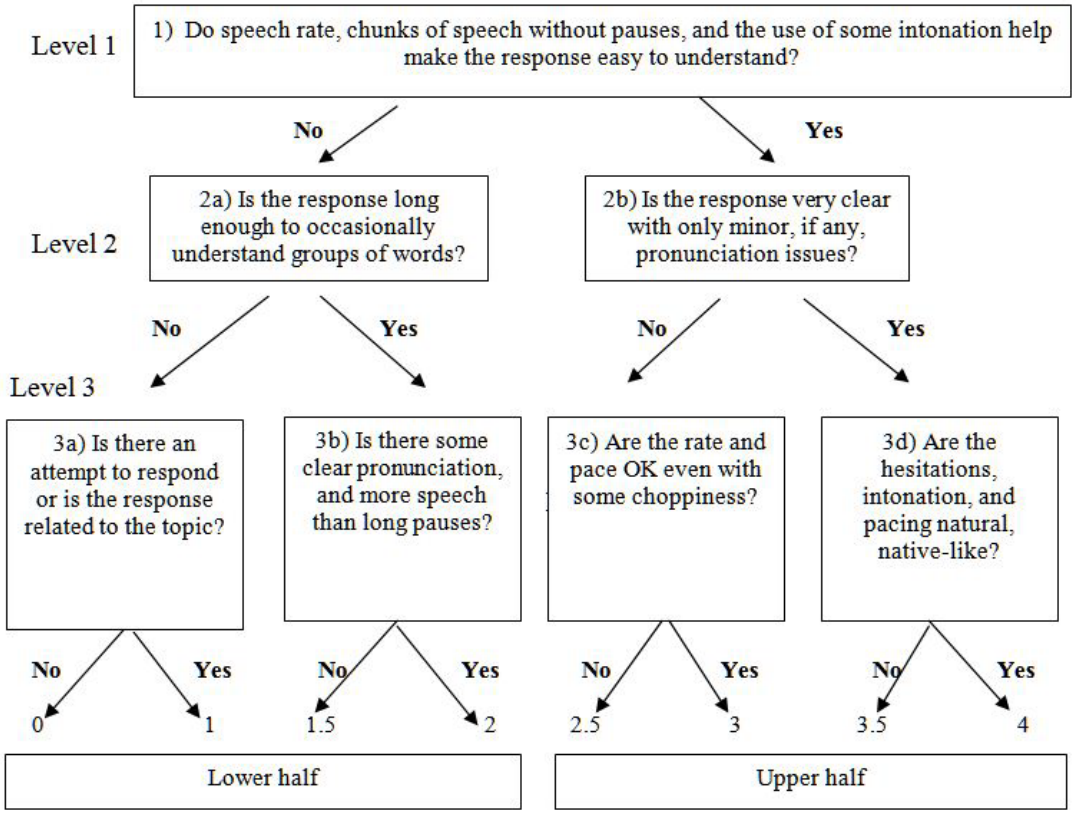


Figure 4. **A proposed rating guide for delivery.**

However, from a pedagogic perspective, the focus on the listener as opposed to the speaker does not provide adequate guidance for improvement. How does a teacher tell a student to improve by affecting another person? While listener effort may be a reasonable criterion for a novice rater, it is not as helpful in instructional settings. Moreover English language teachers may need little effort to understand their students' speech, particularly if they are teaching students of their own language background. The data indicated some differences between the 2–2.5 levels for syllables per second, mean length of run, and overall pitch range, so these were included to focus attention on specific linguistic features. Mean length of run was worded as *chunks of speech without pauses.* Here is the question at the first decision point: Do speech rate, chunks of speech without pauses, and the use of some intonation help make the response easy to understand?

There were two Level 2 decisions. One subdivided the lower half of the responses, between scores of 0–1 and 1.5–2. Both mean length of run and raters' verbal reports indicated that responses at Score Levels 1.5 and 2 had brief stretches in which speech was sustained. By sustained, we mean sequences of words uttered without a noticeable break (similar to Goldman-Eisler's use of *phrase* (1961b); we do not mean the sense of maintaining speech at a consistent level as in the ACTFL guidelines for speaking); these stretches are still very short as compared to high intermediate or advanced speakers. They are sufficiently long though to separate responses from the lower scores of 0 (no attempt to respond) and 1 (very short responses with only occasional words understood). Here is the question at this decision point: Is the response long enough to occasionally understand groups of words?

The other Level 2 decision divided the upper half of the responses between scores of 2.5–3 and 3.5–4. While the phonological variables were not easily divisible at this point, the verbal reports differed in comments about pronunciation. At Score Levels 3 and below, a theme was pronunciation problems associated with accented speech although raters commented that the speech was generally clear. In contrast, raters commented on only minor pronunciation problems at Score Level 3.5 and good pronunciation with little accent at Score Level 4; raters commented that speech was very clear. Here is the question at this decision point: Is the response very clear with only minor, if any, pronunciation issues?

At Level 3, four decision points were needed. To distinguish Score Level 0 from 1, we relied on the verbal protocols and the TOEFL iBT integrated speaking scale, which places responses in which the speaker makes no attempt to respond at 0. Here is the question at this decision point: Is there an attempt to respond or is the response related to the topic?

To distinguish Score Levels 1.5 and 2, we considered the number of silent pauses and pronunciation. Quantitatively the number of silent pauses decreased from Score Level 1.5 to 2, similar to the findings in Brown et al. (2005); also, raters made fewer comments about hesitations and long pauses at Score Level 2 than they did at Score Level 1.5, reflecting trends reported in Kang (2008, 2010) and Kormos and Denes (2004). The other rater theme that distinguished the score levels was poor versus generally clear, adequate pronunciation. The question at this decision point includes these features: Is there some clear pronunciation, and is the response somewhat sustained in a few places? Notice that at some decision points we do have two questions in one. A rater should select yes in the rating guide only when his/her answers to

both questions was yes; for example, yes, there is some clear pronunciation and yes, the response is somewhat sustained in a few places.

Next, to distinguish Score Levels 2.5 and 3, we considered rate (or *pace*, as the raters referred to it) and hesitations. Interestingly, raters used *choppy* and *pace* to refer to rate rather than intonation (cf. Kang, 2008; Kormos & Denes, 2004; Wennerstrom, 2000, 2001). Considering the average rates as measured by the number of syllables per second, there was an increase between Score Levels 2.5 and 3, similar to the distinction Anderson-Hsieh and Venkatagiri (1994) found between intermediate and advanced NNS. Also, at score levels below 2.5 raters commented on inconsistent or choppy pace, whereas pace was not considered to be a theme of raters again until Score Level 4, when they commented on "native-like pace." Here is the question at this decision point: Are the rate and pace OK even with a bit of choppiness?

Finally, to distinguish Score Levels 3.5 and 4, we included hesitations, intonation, and pacing. Raters mentioned that responses at Score Level 3.5 were a bit choppier and had a few more hesitations than they did at Score Level 4. Because several raters and researchers (Brown et al., 2005; Munro & Derwing, 1995) have referred to native-like speech in their descriptions, we have included it in our question. Here is the question at this decision point: Are the hesitations, intonation, and pacing natural, native-like?

Language use. Rather broad categories were used in the questions to distinguish score levels for language use for two main reasons. First, the use of the lexico-grammatical features included in the study was affected by task to a greater degree than the acoustic variables included for delivery (which were largely invariant in reference to task). Second, most raters spoke mainly about vocabulary and grammar in very broad terms. For instance, raters spoke of basic versus sophisticated vocabulary and simple versus complex grammar.

In general, features that were associated with high language use scores were long responses that contained more than basic vocabulary, attempts at complex structures, few errors with basic vocabulary and both simple and complex structures, and automaticity (similar to Brown et al., 2005; Iwashita et al., 2008; Ortega, 1999; Wigglesworth, 1997). Grammatical forms associated with high scores on some tasks were relative clauses, prepositional phrases, passives, adjectives, and adverbs of place and degree (Biber, Gray, & Poonpon, 2011; Norrby & Hakansson, 2007). TTR and the number of lexical verbs were associated with low scores (similar to Brown et al., 2005). Subsequent analysis indicated that responses with very low scores

consisted of very few words that were not repeated, thus yielding a high TTR. This underscores the warning of not computing TTR on samples less than 100 words. These same responses also contained a number of words from the prompt, perhaps accounting for their higher number of lexical verbs. Unexpectedly, in the quantitative results the proportion of high- and low-frequency words was not strongly related to language use scores (cf. Daller et al., 2003; Vermeer, 2000).

The Level 1 decision point between the lower and upper halves (0–2 versus 2.5–4) of the language use scores could have addressed the number of words or errors, or attempts at a wider range of words and more than simple grammatical structures. We decided to use the latter features. First, the raters' comments referred mainly to basic vocabulary and simple sentences at Score Levels 2 and below, but mentioned attempts at more complex language at Score Levels 2.5 and above. Second, as stated above, the quantitative results and previous literature indicated a greater use of more complex forms. Third, attempts at more sophisticated language leads to desirable linguistic goals and rewards effort over accuracy. As shown in Figure 5, the question for this first decision point in the rating guide for language use was as follows: Does the speaker attempt both complex structures and vocabulary?

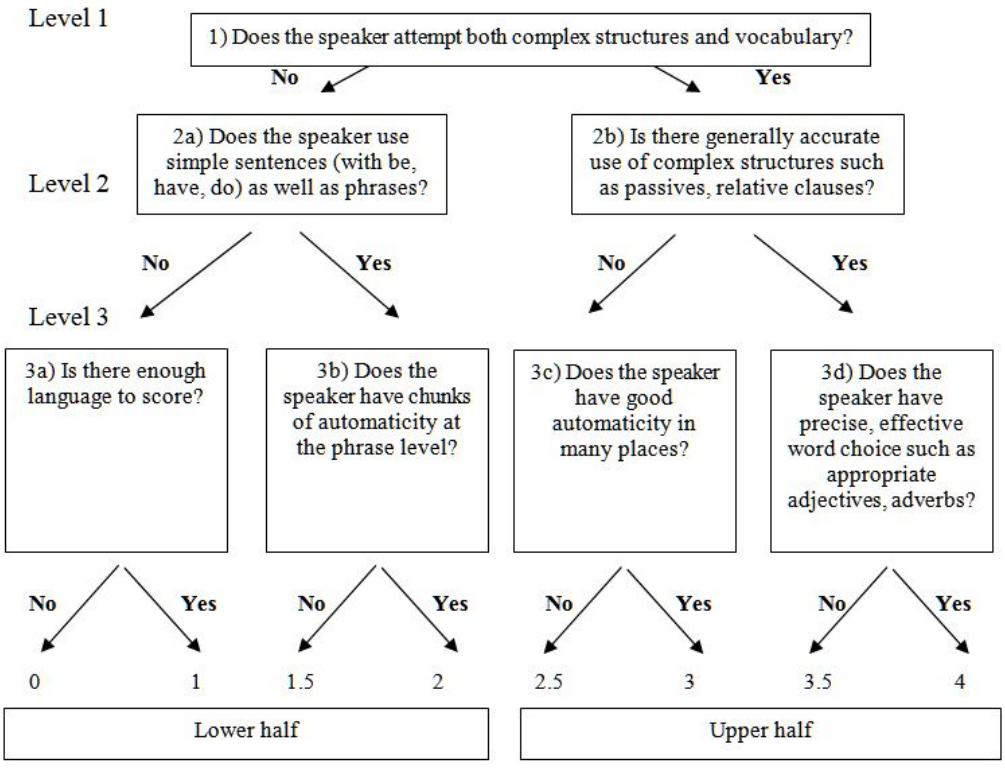


Figure 5. **A proposed rating guide for language use.**

The first Level 2 decision point subdivided the lower half of the responses between 0-1 and 1.5–2. Based on raters' comments there was more frequent use of simple sentences and phrases at the higher score level. Also, the linguistic analysis indicated a higher use of primary verbs (e.g., be, have, do) at Score Levels 1.5–2 than at most other score levels. Here is the question for this decision point: Does the speaker use simple sentences (with be, have, do) as well as phrases?

The second Level 2 decision point subdivided the top half of the responses (2.5–3 versus 3.5–4). The quantitative results indicated that the use of both relative clauses and passives increased with the higher scores. Whereas raters commented on attempts, or some use of complex sentences at Score Level 2.5 and 3, they referred to use of many complex structures with relatively minor errors at the higher score levels, similar to findings by Norrby and Hakansson (2007). Here is the question for this decision point: Is there generally accurate use of complex structures such as passives, relative clauses?

For Level 3, there were four decision points. First, between scores of 0 and 1, we again referred to TOEFL iBT's scale at Score Level 0 as nothing to score and contrasted this with raters' comments of having very little to score. Here is the question for this decision point: Is there enough language to score?

The next decision point was between Score Levels 1.5 and 2. Raters only began to mention automaticity at Score Level 2, saying that speakers had shown some automaticity, albeit at a very low level and usually in small chunks of language or phrases. Here is the question for this decision point: Does the speaker have chunks of automaticity at the phrase level?

Automaticity was also the feature that separated responses at the next decision point—2.5 versus 3. One theme from raters at Score Level 2.5 was that many responses displayed a little automaticity at the phrasal level or with simple sentences, but that overall it was lacking. On the other hand, at Score Level 3, raters reported that good automaticity was found in many places in many responses. From the quantitative analysis, the number of prepositional phrases also increased, but a rater at both score levels commented on errors with prepositions, while no one commented on the use of prepositional phrases. For the time being we have left prepositional phrases out in favor of automaticity. Here is the question for this decision point: Does the speaker have good automaticity in many places?

Finally, the distinction between Score Levels 3.5 and 4 involved vocabulary. At both score levels raters thought that the speakers used a good range of vocabulary, but at Score Level 3.5 they used words such as *fairly good*, *some sophistication*, *some awkwardness*, *mostly appropriate*, whereas at Score Level 4, raters' words were *sophisticated*, *precise*, *effective*. This seems to show both more range and more ability in using appropriate vocabulary. This use of advanced vocabulary supports the notions of Daller et al. (2003) and Vermeer (2000). The quantitative results indicated more use of adverbs of place and degree and adjectives at Score Level 4, but there were so few responses at this level we will use them as examples. Here is the question for this decision point: Does the speaker have precise, effective word choice such as appropriate adjectives, adverbs?

Topic development. Quantitative and qualitative results converged, indicating that conveying the information required by the task in the form of the number of key ideas was the most important feature associated with a high topic development score. However, the inter-coder agreement on the number of key ideas was far from perfect. This indicates that attention should be paid to the presentation of key points and topic notes for raters. Both types of results also showed that high scores were associated with good framing of the response, clear progression of ideas, and a good amount and development of details (as in Brown et al., 2005). Surprisingly, the amount of inaccurate information was not often commented on by raters; it seemed that raters looked for the gist and rewarded detail, but did not penalize for factual errors (cf. Brown et al., 2005). Another interesting finding that bears future monitoring was that raters frequently commented on cohesive responses (as in Brown et al., 2005), which were associated with higher scores, but the linguistic measure of cohesion (reference devices) was negatively associated with high scores (unlike the findings of Ejzenberg, 2000, and Fung & Carter, 2007).

The Level 1 question referred to key ideas. Raters' comments indicated that responses at Score Level 2 and below made vague reference to key points, whereas those at 2.5 and above did fulfill the task by including most, if not all, of the key points. A second feature that was considered was the use of an introduction. The empirical results showed differences at this decision point, but few raters commented on framing or introductions at Score Level 2.5, so this feature was not included here. The first question in Figure 6 was as follows: Are most key ideas included?

The first Level 2 question, separating Score Levels 0–1 from 1.5–2 again uses the notion of task fulfillment, also important in Brown et al. (2005). Raters at the higher score levels thought that speakers did attempt to answer the prompt, albeit at a very superficial level; responses scored lower did not. Here is the question for this decision point: Is the speaker able to respond to prompt at superficial level?

The second Level 2 question aims to separate scores 2.5–3 from 3.5–4. The feature that distinguished these two sets of scores was the progression of ideas. Raters mentioned that responses at the Score Levels 2.5–3 had some problems connecting their ideas, whereas responses at Score Levels 3.5–4 were said to be clear, well connected, and coherent. The empirical results lent some support for this difference as conjunction devices were used somewhat more frequently at the higher levels. Here is the question for this decision point: Is there a clear progression of ideas?

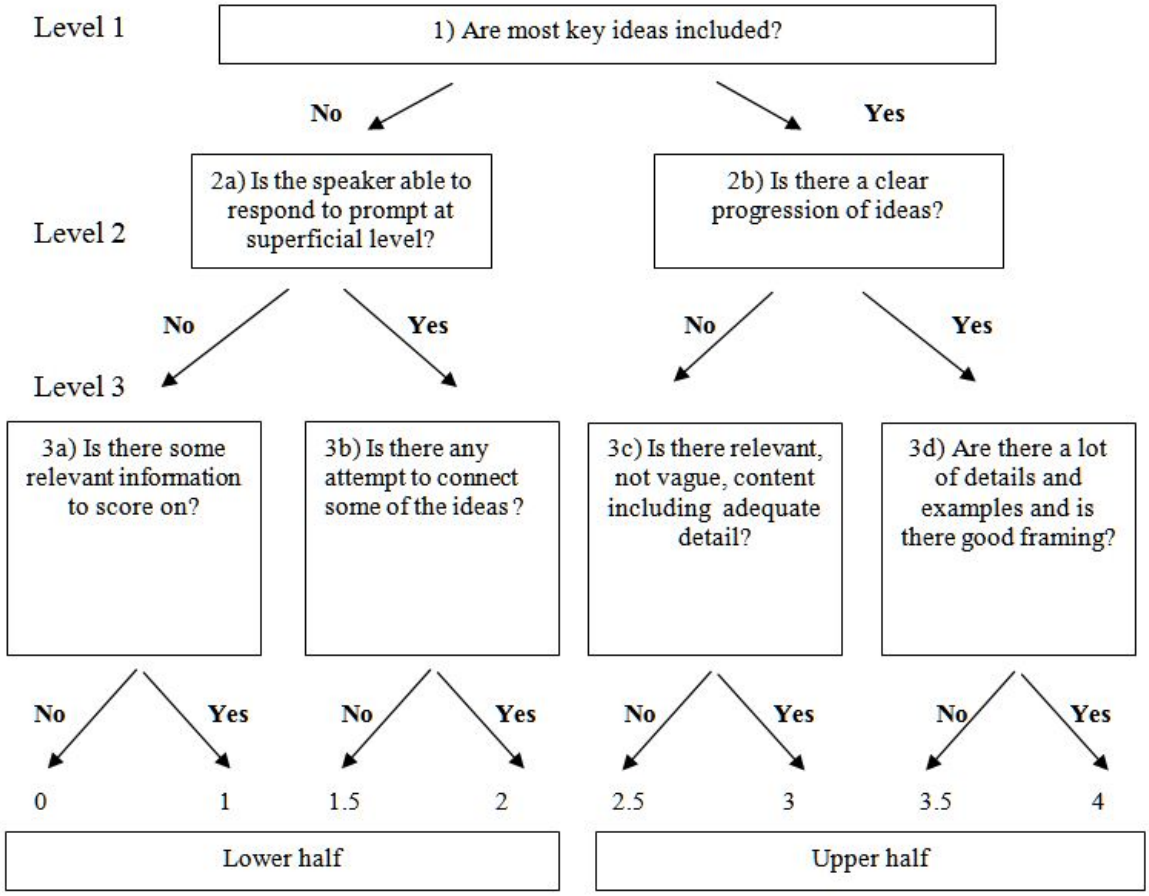


***Figure 6*. A proposed rating guide for topic development.**

The first Level 3 question relied on the TOEFL iBT scale to make a distinction between Score Level 0 and 1—a score of 0 is given if a response was unrelated to the topic or if there was no attempt to respond. Raters commented that responses at Score Level 1 had little relevant information. The question for this decision point includes these ideas: Is there some relevant information to score on?

The second Level 3 question distinguished between Score Levels 1.5 and 2. A distinctive feature was related the connection of ideas. Raters thought that neither score level had clear connections or coherence among ideas, but at Score Level 2 they did mention some use of transitions and coherence that was not done well. We have included the notion of attempt in this question: Is there any attempt to connect some of the ideas?

For the third Level 3 question, raters distinguished Score Levels 2.5 and 3 by the type of detail speakers provided. Responses at Score Level 2.5 were said to fulfill the task but in a vague, incomplete way, with some details but not many. The idea of vagueness was not apparent in the comments given a score of 3; raters also mentioned that these responses had adequate details, but that more details were called for. Here is the question for this decision point: Is there relevant, not vague, content including adequate detail?

Finally, in general, raters did not comment on a very good amount of detail in responses until Score Level 4; also, at this level, raters commented that responses were well framed, similar to Brown et al. (2005). Responses at Score Level 3.5 were said to have an adequate amount of detail, but that more detail could have been given. Also, raters mentioned that framing was inconsistent among responses. Both a good amount of detail and good framing are intended to distinguish these score levels with the following question: Are there a lot of details and examples and is there good framing? Again, we have two questions here. A rater should select yes in the rating guide only when his/her answers to the two questions were both yes.

Discussion

In order to justify three analytic rating guides for delivery, language use, and topic development, three research questions were asked in this study. In this section, we summarize our answers, discuss related issues, and map future directions. First, raters were able to make meaningful distinctions on the expanded scales. Second, while we were successful in identifying linguistic features associated with high and low scores, we were less successful in identifying

linguistic features that distinguished between score levels. Third, and finally, three rating guides have been developed.

At this point, the rating guides represent an expansion of the existing TOEFL iBT rubric for integrated speaking tasks that incorporate findings from previous research, raters' judgments, and speakers' linguistic choices. Their development represents an important, but intermediate, step toward the goal of developing guides to help teachers, students, and novice raters understand which linguistic features are associated with scores. Next steps involve iterations of trialing and revising the proposed rating guides with raters and teachers on a range of speech samples, taking into account three shortcomings in the present study—low inter-rater reliability, some mismatch with the current TOEFL iBT scoring rubric, and the inability to clearly distinguish performance at adjacent score levels.

Trialing and revision are needed to address issues with the raters and the descriptors. The 7-point scales were found to function well, indicating that raters were able to make distinctions between the score levels (Myford, 2006). Anecdotally, during our training sessions more than one rater commented on their desire to assign different scores for a high 3 or a low 3. The reliability of the scores that raters gave was adequate for the purpose of this study, as were the percentages of misfitting raters, but their eventual consistency remains a question. The low percentage of exact agreement between raters, ranging from 34% to 37%, can be examined from three perspectives: lack of training, scale development, and increased number of scale points. First, raters had been given training in using an expanded scale and exemplars were provided as they scored, but they did not receive rigorous training in how to assign scores, as their judgments and thoughts were being elicited for the very purpose of scale development. Therefore, we did not expect or require that inter-rater reliability would be as high as typically found among raters who had been trained to use a standardized scale or scoring scale. Second, lower than desired percentages of exact agreement are not uncommon in studies of scale development in second language speaking and writing when new, innovative approaches are implemented (Knoch, 2007a; Zechner, Higgins, Xi, & Williamson, 2009). Third, increasing the number of rating possibilities from 4 scale points to 7 decreases the likelihood of exact agreement particularly at the midrange of the scale, which is where the majority of speech samples were located in this study (Bunton, Kent, Duffy, Rosenbek, & Kent, 2007; Penny, Johnson, & Gordon, 2000).

Clearly, rater consistency is a standard to expect in an operational setting and, to a lesser degree, in a classroom setting. A better level of consistency must be established.

Improved consistency can be addressed in future studies by better training and also by using clear descriptors. This latter point not only involves consistency, but also relates to the meaningfulness of the scales. Xi and Mollaun (2006) noted that raters thought some of the TOEFL iBT descriptors overlapped. The descriptors need to be understood by raters and teachers, but perhaps more importantly they need to indicate language features that can be directly interpreted in terms of stronger or weaker speaking performance. Are the current descriptors up to the task? Are they in line with the TOEFL iBT rubric?

One potential problem, for example, concerns the inclusion of framing, or having an introduction, as a descriptor for a high score in topic development. As a reviewer pointed out, while shown in studies to be related to scores (e.g., Brown et al., 2005), introductions and conclusions are not required, nor even desired in exemplar responses (particularly as test takers may use large parts of the prompt for framing, thus reducing available response time).

Still another issue regards the use of native speakers or native-like language in the descriptors at the highest level. The use of *native speaker* as a criterion for top-level performance has been, and remains, a controversial issue in language testing. On the one hand, in reality a rating scale has two extreme ends that go on indefinitely, so one doesn't need to be native-like to be in the top category. On the other hand, in reality many of the raters in this study verbalized being native-like as their reason for giving top scores. For the time being, we have included native-like performance in the delivery rating guide. Issues such as these apparent mismatches with TOEFL iBT scoring practice will need to be considered.

The third shortcoming that needs further inquiry involves the specific linguistic features that may distinguish performance at different score levels. Unlike the development of the EBB scales by Upshur and Turner (Turner & Upshur, 2002; Upshur & Turner, 1995, 1999), we did not rely on teachers' consensus to distinguish between score points. Instead, we relied on our interpretation of the data in the study and the literature. Although we were able to adapt their framework of yes/no decisions, we were not able to clearly identify linguistic features that distinguished between two scores. We did find that some features occur with greater or lesser frequency for a given task or for a particular dimension, but these were continuous relationships. When we tried to break down the continuum into categories, we were largely unsuccessful.

Perhaps this was due to the small number of raters included in the study, further reduced by the constraint that they agreed on scores for each task, which masked the role of the distinguishing linguistic features that had been analyzed quantitatively. Was this result due to our methodological approach? Although a number of potential grammatical features were examined, each was counted individually. Another approach to explore this issue would involve examining features in clusters; by grouping grammatical features that have a tendency to co-occur, as Biber has done in his multidimensional analyses (Biber, 1988), we might find that tasks and their score levels can be characterized by several co-occurring features.

At this point, no raters have used the descriptors to make score distinctions. It is our hope that raters will find the analytic rating guides as easy to use and as meaningful as the EBB scales on which they were based. After trialing and revisions, the issues discussed here and many others will need to be resolved. We are hopeful that when the final rating guides are completed teachers and novice raters will be able to make consistent, useful decisions among speakers by following clear descriptors that lead to scores.

There are several steps that are needed to achieve the goals of this project. As alluded to above, trialing of the draft rubrics is needed. One necessary question that must be answered is as follows: Are the three analytic rating guides used consistently by trained ETS raters and ESL/EFL teachers, including both novices and experts? A first step would have a subset of raters from the current study rescore speech samples, using the rating guides. A second step would have novice ETS raters and teachers use the rating guides, still using the speech samples from this study. A subsequent step would add new speech samples using different prompts and for ETS raters using samples from speakers who are planning to take the TOEFL iBT and for the teachers, using samples from speakers who are typical language learners in their schools.

Along with providing background information and their opinions and attitudes about the usefulness of the scoring guides, some raters would be asked to give verbal reports of their experiences. The use of verbal reports was a valuable methodological instrument in the current study (cf. Brown et al., 2005; Green, 1998). Although there was a threat to the interpretation of the scores due to pairing scoring decisions with verbal reports that might affect raters' decisions (Asencion, 2004; Thomson & Isaacs, 2007), this did not turn out to be the case. Information gleaned from future trials will allow us to continually revise and improve the descriptors on the

rating guides. Such information should also help us determine the degree to which raters find the scoring guides useful.

Two other issues should be considered while moving forward. First, for TOEFL iBT raters, a method will need to be devised to convert the three analytic scores to one score based on the current holistic scoring rubric. Such work would allow us to examine the interrelationships among the dimensions. Addressing the assumption that the three dimensions represented subconstructs of speaking, we could examine both convergent and discriminant relationships. Issues regarding the ordering of scoring could also be addressed; for example, is it better to rate one student at a time on all three dimensions, or is it better to rate all speakers on one dimension at a time?

Second, a method will need to be devised to incorporate the use of the rating guide into instructional settings. It is our belief that these rating guides have the potential to provide teachers with tools that are sensitive to students' speaking improvement. How will teachers get access to these rating guides? Will teachers be willing to record all student responses so that they can play them back for three different ratings in delivery, language use, and topic development? Although this is extra work for teachers, will it be offset by positive washback in the form of useful information for both students and teachers?

These questions and the answers we find will lead our way forward. To this point, we have demonstrated how language-testing research can incorporate theoretical and empirical insights to develop instructional tools for new test uses. We believe that there is potential in our work for other scale developers. Building on Turner and Upshur's work (Turner & Upshur, 2002; Upshur & Turner, 1995, 1999), these rating guides represent a novel format for formative assessment with their clear paths for indicating second language speakers' strengths and weaknesses. As one reviewer commented, yes, the procedure is labor intensive, but the potential of extending the research base of high stakes' tests such as TOEFL iBT to instructional settings warrants such an investment.

References

American Council on the Teaching of Foreign Languages. (1999). *ACTFL proficiency guidelines c speaking.* Retrieved from http://www.actfl.org/files/public/Guidelinesspeak.pdf

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: AERA.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, *51*(2) 1–38.

Anderson-Hsieh, J., & Venkatagiri, H. (1994). Syllable duration and pausing in the speech of Chinese ESL speakers. *TESOL Quarterly*, *28*, 807–814.

Asencion, Y. (2004). *Validation of reading-to-write assessment tasks performed by second language learners* (Unpublished doctoral dissertation). Northern Arizona University, Flagstaff.

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

Bachman, L., & Savignon, S. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *The Modern Language Journal*, *70*, 380–390.

Barlow, M. (2000). MonoConc (Version 1.5) [Computer software]. Houston, TX: Athelstan.

Becker, S. (1962). The rating of speeches: Scale independence. *Speech Monographs*, *29*, 38–44.

Becker, S., & Cronkite, G. (1965). Reliability as a function of utilized scale steps. *Speech Teacher*, *14*, 291–293.

Biber, D. (1988). *Variations across speech and writing.* Cambridge, UK: Cambridge University Press.

Biber, D. (2001). *Codes for lexico-grammatical features in program "tag count*." Unpublished manuscript.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., … & Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus* (TOEFL Monograph Series MS-25). Princeton, NJ: Educational Testing Service.

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, *45*, 5–35.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finnegan, E. (1999). *Longman grammar of spoken and written English*. Essex, UK: Pearson Education Limited.

Binghadeer, N. (2008). An acoustic analysis of pitch range in the production of native and nonnative speakers of English. *The Asian EFL Journal Quarterly*, *10*, 96–113.

Boersma, P., & Weenink, D. (2009). Pratt (Version 5.1.17) [Computer software]. Retrieved from http://www.fon.hum.uva.nl/praat/

Breiner-Sanders, K. E., Lowe, P. Jr., Miles, J., & Swender, F. (2000). ACTFL proficiency guidelines—Speaking revised 1999. *Foreign Language Annals*, *33*, 13–18.

Brooks, K. (1957). The construction and testing of a forced choice scale for measuring speaking achievement. *Speech Monographs*, *24*, 65–73.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks* (TOEFL Monograph Series MS-29). Princeton, NJ: Educational Testing Service.

Bunton, K., Kent, R., Duffy, J., Rosenbek, J. & Kent, J. (2007). Listener agreement for auditory-perceptual ratings for Dysarthria. *Journal of Speech*, *Language*, *and Hearing*, *50*, 1481–1495.

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL 2000 speaking framework: A working paper* (TOEFL Monograph Series MS-20). Princeton, NJ: Educational Testing Service.

Bygate, M. (2002). Speaking. In R. B. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (pp. 27–38). Oxford, UK: Oxford University Press.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*, 1–47.

Chapelle, C., Enright, M., & Jamieson, J. (2008). Test score interpretation and use. In C. Chapelle, M. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 1–26). New York, NY: Routledge.

Clark, J., & Clifford, R. (1988). The FSI/ACTFL proficiency scales and testing techniques: Development current status, and needed research. *Studies in Second Language Acquisition*, *10*, 129–147.

Cobb, T. (2002). *The web vocabulary profile*. Retrieved from http://www.lextutor.ca/vp/eng/

Cohen, A. (2013). Verbal reports. In C. Chapelle (Ed.), *Encyclopedia of applied linguistics.* Oxford, UK: John Wiley & Sons.

Colby-Kelly, A., & Turner, C. (2007). AFL research in the L2 classroom and evidence of usefulness: Taking formative assessment to the next level. *The Canadian Modern Language Review*, *64*, 9–37.

Couper-Kuhlen, E. (1996). The prosody of repetition: On quoting and mimicry. In E. Couper-Kuhlen & M. Selting (Eds.), *Prosody in conversation: Interactional studies* (pp. 366–405). Cambridge, UK: Cambridge University Press.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213–238.

Creswell, J. (2009). *Research design: Qualitative*, *quantitative*, *and mixed methods approaches.* Thousand Oaks, CA: SAGE.

Creswell, J., & Plano Clark, V. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: SAGE.

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Crystal, D. (1991). *A dictionary of linguistics and phonetics.* Oxford, UK: Basil Blackwell.

Cumming, A., Kantor, R., & Powers, D. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework* (TOEFL Monograph Series MS-22). Princeton, NJ: Educational Testing Service.

Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, *24*, 197–222.

Dandonoli, P., & Henning, G. (1990). An investigation of the construct validity of the ACTFL proficiency guidelines and oral interview procedure. *Foreign Language Annals*, *23*, 131–151.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing.* Cambridge, UK: Cambridge University Press.

Derwing, T., & Munro, M. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. S*tudies in Second Language Acquisition*, *20*, 1–16.

Derwing, T., & Munro, M. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, *39*, 379–397.

Derwing, T., Rossiter, M., Munro, M., & Thomson, R. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, *54*, 655–679.

Douglas, D., & Smith, J. (1997). *Theoretical underpinnings of the test of spoken English revision project.* (TOEFL Monograph Series MS-9). Princeton, NJ: Educational Testing Service.

Edwards, J., & Lampert, M. (Eds.). (1993). *Talking data: Transcription and coding in discourse research*. Hillsdale, NJ: Erlbaum.

Elluminate, Inc. (2008). Video conferencing [Computer software]. Retrieved from http://www.elluminate.com/

Embretson, S. (1983). Construct validity: Construct representation vs. the nomothetic span. *Psychological Bulletin*, *93*, 179–187.

Enright, M., Bridgeman, B., Eignor, D., Kantor, R., Mollaun, P., Nissan, S., Powers, D., & Schedl, M. (2008). Prototyping new assessment tasks. In C. Chapelle, M. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 97–144). New York, NY: Routledge.

Ejzenberg, R. (2000). The juggling act of oral proficiency: A psycho-sociolinguistic metaphor. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 287–313). Ann Arbor: University of Michigan Press.

Ericsson, K. A., & Simon, H. (1984). *Protocol analysis. Verbal reports as data.* Cambridge, MA: The MIT Press.

ETS. (2007a). *Test and score data summary for TOEFL computer-based and paper-based tests: July 2005–June 2006 test data.* Princeton, NJ: Author. Retrieved from http://www.ets.org/Media/Research/pdf/TOEFL-SUM-0506-CBT.pdf

ETS. (2007b). *Test and score data summary for TOEFL Internet-based test: September 2005–December 2005 test data.* Princeton, NJ: Author. Retrieved from http://www.ets.org/Media/Research/pdf/TOEFL-SUM-0506-iBT.pdf

ETS. (2008). *TOEFL iBT test—Integrated speaking rubrics (scoring standards).* Princeton, NJ: Author. Retrieved from http://www.ets.org/Media/Tests/TOEFL/pdf/Integrated_Speaking_Rubrics_2008.pdf

Foster, P., & Skehan, P. (1996). The influence of planning and performance in task based learning. *Studies in Second Language Acquisition*, *18*, 299–324.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*, 354–375.

Freed, B. (2000). Is fluency, like beauty, in the eyes (and ears) of the beholder? In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 243–265). Ann Arbor: The University of Michigan Press.

Fulcher, G. (1987). Tests of oral performance: The need for data-based criteria. *ELT Journal*, *41*, 287–291.

Fulcher, G. (1993). *The construct validation of rating scales for oral tests in English as a foreign language*. Unpublished PhD thesis, University of Lancaster, UK.

Fulcher, G. (1996a). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, *13*, 208–238.

Fulcher, G. (1996b). Invalidating validity claims for the ACTFL oral rating scales. *System*, *24*, 163–172.

Fulcher, G. (1997). The testing of speaking in a second language. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education* (pp. 75–85). Norwell, MA: Kluwer Academic Publishers.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, *28*, 5–29.

Fung, L., & Carter, R. (2007). Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied Linguistics*, *28*, 410–439.

Goldman-Eisler, F. (1958). The predictability of words in context and the length of pauses in speech. *Language and Speech*, *1*, 226–231.

Goldman-Eisler, F. (1961a). Continuity of speech utterance, its determinants and its significance. *Language and Speech*, *4*, 220–231.

Goldman-Eisler, F. (1961b). The distribution of pause duration in speech. *Language and Speech*, *4*, 232–237.

Goldman-Eisler, F. (1961c). The significance of changes in the rate of articulation. *Language and Speech*, *4*, 171–174.

Green, A. (1998). *Verbal protocol analysis in language testing research. A handbook.* Cambridge, UK: Cambridge University Press.

Greene, J., Caracelli, V., & Graham, W. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, *11*, 255–274.

Gulliksen, H. (1950). Intrinsic validity. *American Psychologist*, *5*, 511–517.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London, UK: Longman.

Hincks, R. (2005). Measures and perceptions of liveliness in student oral presentation speech: A proposal for an automatic feedback mechanism. *System*, *33*, 575–591.

Hofmans, J., Theuns, P., & Mairesse, O. (2007). Impact of the number of response categories on linearity and sensitivity of self-anchoring scales. *Methodology*, *3*, 160–169.

Hudson, T. (2005). Trends in assessment scales and criterion-referenced language assessment. *Annual Review of Applied Linguistics*, *25*, 205–227.

Hunt, K. (1966). Recent measures in syntactic development. *Elementary English*, *43*, 732–739.

Inoue, M. (2009). Health sciences communication skills test: The development of a rating scale. *Melbourne Papers in Language Testing*, *14*, 55–90.

Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English-speaking graduate students. *The Canadian Modern Language Review*, *64*, 555–580.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*, 24–49.

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning*, *51*, 401–436.

Jiang, N. (2007). Selective integration of linguistic knowledge in adult second language learning. *Language Learning*, *57*, 1–33.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed method research: A research paradigm whose time has come. *Educational Researcher*, *33*, 7, 14–26.

Jourdenais, R. (2001). Cognition, instruction, and protocol analysis. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 354–375). Cambridge, UK: Cambridge University Press.

Juffs, A. (1990). Tone, syllable structure and interlanguage phonology: Chinese learners' stress errors. *International Review of Applied Linguistics*, *28*, 99–117.

Kang, O. (2008). Ratings of L2 performance in English: Relative impact of rater characteristics and acoustic measures of accentedness. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *6*, 181–205.

Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System*, *38*, 301–315.

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal*, *94*. 554–566.

Kim, D. (2013). Mixed methods. In C. Chapelle (Ed.), *Encyclopedia of applied linguistics.* Oxford, UK: John Wiley & Sons.

Knoch, U. (2007a). Do empirically developed rating scales function differently to conventional rating scales for academic writing? *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *5*, 1–36.

Knoch, U. (2007b). "Little coherence, considerable strain for reader": A comparison between two rating scales for the assessment of coherence. *Assessing Writing*, *12*, 108–128.

Koh Chin, H., & Bhandari, R. (2006). *Open doors 2006: Report on international education exchange.* New York, NY: Institute of International Education.

Koponen, M., & Riggenbach, H. (2000). Overview: Varying perspectives on fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 5–24). Ann Arbor: University of Michigan Press.

Kormos, J., & Denes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*, 145–164.

Ladefoged, P. (2006). *A course in phonetics.* Boston, MA: Thomson Wadsworth.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*, 387–417.

Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 25–42). Ann Arbor: University of Michigan Press.

Liskin-Gasparro, J. (1984). The ACTFL proficiency guidelines: A historical perspective. In T. Higgs (Ed.), *Teaching for proficiency*, *the organizing principle* (pp. 11–42). Lincolnwood, IL: National Textbook Company.

Lowe, P., Jr. (1986). Proficiency: Panacea, framework, process? A reply to Kramsch, Schulz, and, particularly, to Bachman and Savignon. *The Modern Language Journal*, *70*, 391–397.

Luoma, S. (2004). *Assessing speaking*. Cambridge, UK: Cambridge University Press.

Mackenzie, A. S. (2005, October). *Current developments in EFL curriculum reform in Thailand*. Paper presented at the 7th International Language and Development conference, Addis Ababa, Ethiopia.

McNamara, T. (1996). *Measuring second language performance.* New York, NY: Longman.

McNamara, T. (1997). Performance testing. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education: Vol. 7. Language testing and assessment* (pp. 131–139). Norwell, MA: Kluwer Academic.

McNamara, T. (2000). *Language testing*. Oxford, UK: Oxford University Press.

Mennen, I. (1998). Second language acquisition of intonation: The case of peak alignment. *Chicago Linguistic Society*, *34*, 327–341.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*, 1012–1027.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13–23.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded source book*. Thousand Oaks, CA: Sage.

Miller, D., Linn, R., & Gronlund, N. (2009). *Measurement and assessment in teaching* (10th ed.). Upper Saddle River, NJ: Pearson.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, *63*, 81–97.

Munro, M., & Derwing, T. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*, 73–97.

Munro, M., & Derwing, T. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech. *Studies in Second Language Acquisition*, *23*, 451–468.

Myford, C. (2006). *Analyzing rating data using Linacre's Facets computer program: A set of training materials to learn to run the program and interpret output*. Unpublished manuscript.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-faceted Rasch measurement: Part I. *Journal of Applied Measurement*, *4*, 386–422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-faceted Rasch measurement: Part II. *Journal of Applied Measurement*, *5*, 189–227.

Nakatsuhara, F. (2007). Developing a rating scale to assess English speaking skills of Japanese upper-secondary students. *Essex Graduate Student Papers in Language and Linguistics*, *9*, 83–103.

Norrby, C., & Hakansson, G. (2007). The interaction of complexity and grammatical processability: The case of Swedish as a foreign language. *International Review of Applied Linguistics*, *45*, 45–68.

North, B. (1993). *The development of descriptors on scales of language proficiency* (NFLC Occasional Papers). Washington, DC: Johns Hopkins University.

North, B. (1995). The development of common framework scale of descriptors of language proficiency based on a theory of measurement. *System*, *23*, 445–465.

North, B. (2000). *The development of a common framework scale of language proficiency*. New York, NY: Peter Lang.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing*, *15*, 217–263.

Onwuegbuzie, A. J. (2007). Mixed methods research in sociology and beyond. In G. Ritzer (Ed.), *The Blackwell encyclopedia of sociology* (Vol. 6, pp. 2978–2981). Oxford, UK: Blackwell.

Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition*, *21*, 109–148.

Penny, J., Johnson, R., & Gordon, B. (2000). Using rating augmentation to expand the scale of an analytic rubric. *The Journal of Experimental Education*, *68*, 269–287.

Phillips, N., Segalowitz, N., O'Brien, I., & Yamasaki, N. (2004). Semantic priming in a first and second language: Evidence from reaction time variability and event-related brain potentials. *Journal of Neurolinguistics*, *17*, 237–263.

Pickering, L. (2004). The structure and function of intonational paragraphs in native and nonnative speaker instructional discourse. *English for Specific Purposes*, *23*(1), 19–43.

Poonpon, K. (2009). *Expanding a second language speaking rating scale for instructional and assessment purposes* (Unpublished doctoral dissertation). Northern Arizona University, Flagstaff.

Preston, C., & Colman, A. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*, 1–15.

Rea-Dickins, P. (2004). Understanding teachers as agents of assessment. *Language Testing*, *21*, 249–258.

Richards, J., Platt, J., & Platt, H. (1992). *Dictionary of language teaching & applied linguistics.* Essex, UK: Longman.

Rimmer, W. (2006). Measuring grammatical complexity: The Gordian knot. *Language Testing*, *23*, 497–519.

Rosenfeld, M., Leung, S., & Oltman, P. (2001). *The reading*, *writing*, *speaking*, *and listening tasks important for academic success at the undergraduate and graduate levels* (TOEFL Monograph MS-21). Princeton, NJ: ETS.

Sakui, K. (2004). Wearing two pairs of shoes: Language teaching in Japan. *ELT Journal*, *58*, 155–163.

Savignon, S. J. (1985). Evaluation of communicative competence: The ACTFL provisional proficiency guidelines. *The Modern Language Journal*, *69*, 129–134.

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing*, *24*, 355–390.

Segalowitz, N. (2003). Automaticity and second languages. In C. Doughty & M. Long (Eds.), *The handbook of second language acquisition* (pp. 382–408). Oxford, UK: Blackwell Publishing.

Shohamy, E. (2006). *Language policy: Hidden agendas and new approaches*. New York, NY: Routledge.

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, *49*, 93–120.

Solórzano, H. (2006). *NorthStar: Building skills for the TOEFL iBT*. White Plains, NY: Pearson Longman.

Stansfield, C., & Kenyon, D. (1992). The development and validation of a simulated oral proficiency interview. *The Modern Language Journal*, *76*, 130–141.

Taniguchi, M. (2001, April). *Japanese EFL learners' weak points in English intonation.* Paper presented at the 2001 Phonetics Teaching and Learning conference, London, UK. Retrieved from http://www.phon.ucl.ac.uk/home/johnm/ptlc2001/pdf/tani2.pdf

Tanskanen, S. (2004). Cohesion and collaboration: Patterns of cohesion in spoken and written dialogue. In K. Aijmer (Ed.), *Discourse patterns in spoken and written corpora* (pp. 89–110). Philadelphia, PA: John Benjamins.

Taylor, C., & Angelis, P. (2008). The evolution of TOEFL. In C. Chapelle, M. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 27–54). New York, NY: Routledge.

Thomson, R., & Isaacs, T. (2008, March–April). *What lies beneath: Influences of rater assessments of L2 pronunciation.* Paper presented at the 2008 American Association of Applied Linguistics conference, Washington, DC.

Timkova, R. (2001, April). *Intonation of English in the process of second language acquisition*. Paper presented at the 2001 Phonetics Teaching and Learning conference, London, UK.

Toivanen, J., Väyrynen, E., & Seppänen, T. (2004). Automatic discrimination of emotion from spoken Finnish. *Language and Speech*, *47*, 383–412.

Turner, C., & Upshur, J. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly*, *36*, 49–70.

Upshur, J., & Turner, C. (1995). Constructing rating scales for second language tests. *ELT Journal*, *49*, 3–12.

Upshur, J., & Turner, C. (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing*, *16*, 82–111.

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, *17*, 65–83.

Wall, D., & Horák, T. (2006). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 1*, *the baseline study* (TOEFL Monograph Series MS-34). Princeton, NJ: Educational Testing Service.

Wall, D., & Horák, T. (2008). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 2*, *coping with change* (TOEFL iBT Research Series No. 5). Princeton, NJ: Educational Testing Service.

Wennerstrom, A. (1994). Intonational meaning in English discourse: A study of nonnative speakers. *Applied Linguistics*, *15*, 399–421.

Wennerstrom, A. (1998). Intonation as cohesion in academic discourse: A study of Chinese speakers of English. *Studies in Second Language Acquisition*, *42*, 1–13.

Wennerstrom, A. (2000). The role of intonation in second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 102–127). Ann Arbor: University of Michigan Press.

Wennerstrom, A. (2001). *The music of everyday speech.* Oxford, UK: Oxford University Press.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing*, *14*, 85–106.

Xi, X., & Mollaun, P. (2006). *Investigating the utility of analytic scoring for the TOEFL Academic Speaking Test (TAST*; TOEFL iBT Research Series No. 1). Princeton, NJ: Educational Testing Service.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, *51*, 883–895.

List of Appendices

Appendix A

List of Reference and Conjunction Devices

Types of cohesive devices	Subcategories	Words/phrases
Reference	Personal	I, me, mine, my, you, your, yours, we, us, our, ours, he, him, his, she, her, hers, it, its, they, them, their, theirs
	Demonstrative	this/these, that/those, the, here, there, then
	Comparative	same, similar(ly), such, more, less, as many, as + adjectives, comparatives, superlatives
Conjunction	Addition	first(ly), second(ly), third(ly) (etc.), in the first/ second place, first of all, for one thing, for another thing, to begin with, next, and, and also, or, nor, in addition, additionally, further, furthermore, moreover, add to that, finally, lastly, in the end
	Apposition	for example, for instance, e.g., in other words, that is
	Result	therefore, consequently, as a result, thus, so
	Contrast or adversative	on the other hand, in contrast, alternatively, however, yet, but, though, although, even though, conversely, on the contrary, instead
	Transition	incidentally, in the meantime, meanwhile, for the moment, by the way
	Summation	in sum, to conclude, in conclusion

Appendix B

Descriptive Statistics of Linguistic Variables

Delivery

Four acoustic variables previously shown to be statistically significant (syllables per second, mean length of run, number of silent pauses, and overall pitch range) were analyzed. As shown in Table B1, the number of syllables per second averaged 2.82, mean length of run averaged 3.67, the average number of silent pauses was 49.25, and the overall pitch range averaged 5.74. Examination of histograms indicated that all four variables were normally distributed.

Table B1

Descriptive Statistics for Delivery Measures

Variable	*N*	M	SD
Syllables per second	231	2.82	.69
Mean length of run	231	3.67	1.16
Number of silent pauses	231	49.25	11.92
Overall pitch range	226	5.74	2.27

Note. Due to poor sound quality, five responses could not be analyzed for pitch.

Language Use

Two different sets of variables were examined. The first set included variables that had been found important in previous research. These *literature-driven* variables consisted of TTR, low versus high frequency words, error-free C-units, complement clauses, adverbial clauses, relative clauses, prepositional phrases, stance adverbs, and passives. The second set of variables was what we called *data-driven*. These were variables from about 70 lexico-grammatical categories tagged by Biber's computer program (2001). The counts of all of these features were normed to 100 words so that responses of different lengths could be compared. Based on preliminary correlational analysis with the average language use score for both sets of variables, 12 variables emerged. The descriptive statistics for these language use variables are displayed in Table B2. Inspection of histograms, Zskewness, and Zkurtotsis revealed that the distributions for relative clauses, passives, adverbs of place, and adverbs of degree were not normally distributed;

LG10 transformations were unsuccessful in normalizing these distributions for subsequent statistical analyses.

Table B2

Descriptive Statistics for Language Use Measures

Variable	*N*	M	SD
Type-token ratio	229	.55	.08
Error-free C-units	231	.31	.22
Relative clauses	231	.67	.82
Prepositional phrases	230	7.40	2.40
Wh-words	231	.12	.33
Passives	231	.62	.88
Adverbs of place	230	.48	.83
Word count	231	123.75	30.65
Adjectives	230	3.55	2.13
Lexical verbs	229	3.83	2.23
Primary verbs	229	1.06	1.05
Adverbs of degree	230	1.03	1.21

Topic Development

Four variables were analyzed in relation to topic development—key ideas, conjunction devices, reference devices, and introduction. Descriptive statistics for three of the variables are displayed in Table B3. The distributions of conjunction devices were not normal. Of the 231 responses, 124 (54%) had an introduction that framed the response.

Table B3

Descriptive Statistics for Topic Development Measures

Variable	*N*	M	SD
Key ideas	231	1.59	.93
Conjunction devices	231	1.26	1.29
Reference devices	231	8.61	3.90

Appendix C

Measures by Score Levels and Tasks

Table C1

Mean and Standard Deviation on Delivery Measures by Score Level and Task

Delivery score	Task	*N*	Syllables per second	Mean length of run	Number of silent pauses	Overall pitch range
1	3	0	[a]	[a]	[a]	[a]
1	4	0	[a]	[a]	[a]	[a]
1	5	1	[a]	[a]	[a]	[a]
1	6	0	[a]	[a]	[a]	[a]
1.5	3	4	1.38 (.58)	3.35 (2.30)	36.25 (21.36)	3.83 (1.81)
1.5	4	3	2.07 (.47)	2.25 (.26)	57.33 (18.61)	2.76 (1.02)
1.5	5	0	[a]	[a]	[a]	[a]
1.5	6	5	2.33 (.50)	2.49 (.12)	57.80 (14.52)	4.11 (1.72)
2	3	6	2.44 (.54)	2.89 (.49)	53.17 (15.33)	3.91 (1.82)
2	4	5	2.67 (.50)	2.93 (.93)	57.75 (10.50)	4.32 (2.34)
2	5	9	2.18 (.71)	2.58 (.34)	51.00 (12.13)	4.37 (1.48)
2	6	6	2.31 (.22)	2.70 (.33)	52.75 (6.60)	5.08 (1.71)
2.5	3	9	3.00 (.53)	3.41 (.86)	55.44 (10.17)	5.31 (2.58)
2.5	4	8	2.78 (.30)	3.26 (.76)	54.00 (10.69)	4.93 (1.97)
2.5	5	7	2.75 (.64)	3.05 (.24)	54.86 (10.64)	4.70 (1.40)
2.5	6	5	2.95 (.57)	3.31 (1.29)	59.60 (18.26)	6.51 (2.07)
3	3	5	2.72 (.70)	4.40 (.94)	39.20 (11.78)	6.98 (.96)

Delivery score	Task	*N*	Syllables per second	Mean length of run	Number of silent pauses	Overall pitch range
3	4	9	3.25 (.45)	3.98 (.73)	51.22 (10.28)	7.60 (1.75)
3	5	8	3.32 (.48)	4.70 (1.18)	45.38 (11.11)	6.56 (1.43)
3	6	6	3.09 (.47)	4.09 (1.11)	47.83 (7.33)	7.79 (2.04)
3.5	3	6	3.24 (.46)	4.72 (.92)	43.17 (9.54)	7.09 (1.48)
3.5	4	2	3.91 (.70)	5.50 (.52)	43.50 (3.54)	10.20 (.72)
3.5	5	4	2.99 (.26)	4.88 (1.07)	38.75 (7.59)	6.62 (1.53)
3.5	6	3	3.48 (.25)	4.49 (.67)	48.67 (10.97)	8.08 (1.88)
4	3	1	[a]	[a]	[a]	[a]
4	4	2	3.68 (.25)	4.76 (1.02)	48.50 (13.44)	9.27 (1.46)
4	5	0	[a]	[a]	[a]	[a]
4	6	1	[a]	[a]	[a]	[a]

Note. Only scores on which raters agreed are included. SD in parentheses. $N = 115$.

[a] One or fewer scores at that level for task.

Table C2

Mean and Standard Deviation on Language Use Measures by Score Level and Task

Score	Task	*n*	Type-token ratio	Error-free C-unit	Relative clauses	Prepositional phrases	Passive	Place adverb	Word count	Adjectives	Lexical verbs	Primary verbs	Degree adverb
1	3	1	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]
1	4	1	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]
1	5	0	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]
1	6	1	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]
1.5	3	5	.60 (.05)	.00 (.00)	.82 (1.44)	8.82 (3.31)	.62 (.85)	.15 (.34)	78.80 (31.70)	4.19 (2.96)	6.47 (2.12)	.93 (1.36)	.29 (.64)
1.5	4	3	.61 (.13)	.19 (.07)	.40 (.69)	4.90 (1.88)	.00 (.00)	.00 (.00)	84.67 (12.01)	4.90 (3.09)	5.63 (2.62)	1.59 (.70)	[a]
1.5	5	3	.57 (.06)	.15 (.13)	.36 (.63)	4.65 (.84)	.00 (.00)	1.04 (1.09)	83.33 (27.06)	1.26 (1.09)	5.82 (2.26)	.63 (1.09)	[a]
1.5	6	3	.45 (.12)	.14 (.17)	1.07 (.53)	3.74 (1.88)	.23 (.40)	1.89 (2.68)	108.00 (43.00)	4.07 (3.73)	3.29 (1.62)	3.47 (1.46)	1.11 (.96)
2	3	2	.51 (.11)	.05 (.07)	.38 (.54)	7.73 (3.37)	.00 (.00)	1.13 (1.59)	110.00 (29.70)	3.77 (1.02)	3.39 (1.56)	2.47 (1.90)	.56 (.79)
2	4	4	.55 (.11)	.22 (.16)	.91 (.77)	6.74 (1.56)	.98 (1.15)	.77 (.90)	129.75 (38.68)	4.00 (2.78)	2.05 (1.79)	.94 (.89)	.88 (.79)
2	5	4	.56 (.10)	.28 (.24)	.75 (.59)	5.67 (2.95)	.79 (.59)	.43 (.50)	110.75 (29.97)	1.44 (1.10)	2.37 (.65)	1.15 (1.21)	.74 (1.00)

Score	Task	*n*	Type-token ratio	Error-free C-unit	Relative clauses	Prepositional phrases	Passive	Place adverb	Word count	Adjectives	Lexical verbs	Primary verbs	Degree adverb
2	6	4	.57 (.11)	.13 (.15)	.48 (.96)	6.60 (1.13)	1.17 (1.21)	.45 (.52)	104.75 (25.16)	2.32 (1.97)	5.18 (2.48)	1.50 (2.03)	1.79 (1.43)
2.5	3	5	.58 (.04)	.30 (.14)	.17 (.38)	8.45 (2.02)	1.40 (1.68)	.15 (.34)	121.20 (15.91)	5.22 (3.14)	2.24 (1.27)	.34 (.47)	.63 (.61)
2.5	4	2	.64 (.03)	.00 (.00)	1.82 (.74)	8.06 (1.46)	1.95 (1.66)	.00 (.00)	102.50 (36.06)	5.07 (3.85)	2.86 (.37)	1.17 (1.65)	.78 (1.10)
2.5	5	7	.53 (.08)	.40 (.10)	.25 (.43)	5.83 (1.38)	.34 (.91)	.00 (.00)	149.71 (29.04)	4.20 (1.81)	3.20 (1.09)	1.37 (1.51)	1.47 (1.27)
2.5	6	2	.50 (.06)	.33 (.00)	.38 (.54)	3.68 (.89)	.81 (.07)	.00 (.00)	123.50 (10.61)	1.96 (1.55)	5.31 (1.03)	1.20 (.47)	.43 (.61)
3	3	4	.54 (.07)	.23 (.18)	1.71 (1.24)	8.23 (1.86)	.72 (.10)	.34 (.40)	141.00 (19.54)	3.39 (.90)	3.89 (1.88)	1.12 (.77)	.53 (.70)
3	4	4	.57 (.09)	.51 (.15)	.55 (.38)	10.35 (.96)	.52 (.66)	.00 (.00)	140.25 (14.57)	4.41 (1.36)	2.07 (1.53)	.93 (.75)	1.53 (1.50)
3	5	7	.52 (.03)	.32 (.12)	.48 (.50)	7.63 (1.99)	.30 (.38)	.18 (.32)	146.00 (16.89)	3.99 (1.10)	2.83 (1.47)	1.78 (.56)	.58 (.47)
3	6	2	.48 (.01)	.34 (.01)	.36 (.50)	10.08 (.79)	.00 (.00)	.95 (.34)	154.50 (19.91)	2.50 (1.52)	5.93 (1.64)	.35 (.50)	1.25 (.76)
3.5	3	6	.55 (.05)	.50 (.39)	1.09 (1.02)	9.23 (2.19)	1.18 (.96)	.34 (.38)	130.67 (36.19)	3.55 (1.34)	2.67 (1.63)	.57 (.70)	1.78 (2.98)
3.5	4	4	.57 (.03)	.32 (.08)	1.37 (1.13)	8.65 (.27)	1.22 (1.01)	.36 (.42)	141.50 (9.26)	6.82 (3.38)	2.11 (.77)	1.37 (.98)	1.03 (.90)

Score	Task	*n*	Type-token ratio	Error-free C-unit	Relative clauses	Prepositional phrases	Passive	Place adverb	Word count	Adjectives	Lexical verbs	Primary verbs	Degree adverb
3.5	5	4	.53 (.06)	.57 (.26)	1.03 (.71)	6.40 (1.60)	.60 (.70)	.84 (1.02)	159.25 (23.00)	2.32 (1.27)	2.36 (.48)	.66 (.60)	.62 (.52)
3.5	6	2	.52 (.09)	.38 (.09)	1.14 (.59)	5.68 (2.93)	1.89 (.63)	1.14 (.59)	134.00 (7.07)	2.27 (1.17)	5.27 (1.64)	.36 (.51)	.75 (.04)
4	3	0	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]
4	4	0	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]
4	5	0	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]	[a]
4	6	3	.51 (.02)	.71 (.10)	.20 (.35)	7.32 (1.73)	.87 (.92)	2.37 (1.18)	153.67 (21.46)	2.84 (.40)	6.48 (1.75)	.81 (.93)	2.63 (1.83)

Note. Only scores on which raters agreed are included. SD in parentheses. $N = 115$.

[a]One or fewer scores at that level.

Table C3

Mean and Standard Deviation on Topic Development Measures by Score Level and Task

Score	Task	*n*	Key ideas	Linking adverbials	Reference markers	Introduction
1	3	2	[a]	[a]	13.65 (10.99)	[a]
1	4	3	.17 (.29)	1.44 (1.81)	7.39 (4.38)	.33 (.58)
1	5	1	[a]	[a]	[a]	[a]
1	6	2	[a]	1.67 (2.36)	15.12 (4.88)	.50 (.71)
1.5	3	2	.50 (.71)	1.69 (.23)	4.44 (2.35)	.50 (.71)
1.5	4	8	.25 (.38)	1.21 (1.13)	9.38 (5.01)	.38 (.52)
1.5	5	3	1.67 (.28)	.69 (.72)	11.00 (4.21)	[a]
1.5	6	2	.25 (.35)	.81 (1.14)	6.80 (.49)	[a]
2	3	6	1.00 (.71)	1.38 (1.83)	9.11 (4.77)	.17 (.41)
2	4	4	.75 (.87)	1.55 (1.40)	10.96 (4.51)	.50 (.58)
2	5	4	1.75 (.50)	1.90 (1.31)	8.27 (2.42)	[a]
2	6	4	1.25 (.87)	.95 (1.15)	7.81 (2.86)	.50 (.58)
2.5	3	2	1.75 (.42)	1.06 (.90)	4.08 (2.46)	.67 (.52)
2.5	4	5	.70 (.76)	.27 (.37)	7.12 (2.76)	.80 (.45)
2.5	5	5	1.70	1.97	9.47	[a]

Score	Task	*n*	Key ideas	Linking adverbials	Reference markers	Introduction
			(.45)	(2.74)	(3.74)	
2.5	6	10	1.75 (.89)	1.37 (1.05)	8.51 (4.03)	.70 (.48)
3	3	3	1.67 (.29)	1.70 (.75)	9.57 (2.92)	.67 (.58)
3	4	5	1.50 (.73)	.76 (.94)	7.05 (3.08)	.80 (.45)
3	5	8	2.13 (.52)	.73 (.76)	9.49 (3.36)	.38 (.52)
3	6	5	2.80 (.27)	2.02 (1.30)	7.35 (3.36)	[a]
3.5	3	6	1.92 (.58)	3.19 (1.78)	10.17 (4.31)	.83 (.41)
3.5	4	7	1.79 (.39)	1.15 (.74)	8.08 (2.34)	.71 (.49)
3.5	5	4	2.50 (.58)	1.38 (1.10)	7.65 (3.21)	.75 (.50)
3.5	6	6	2.58 (.38)	1.24 (.91)	6.04 (2.62)	[a]
4	3	1	[a]	[a]	[a]	[a]
4	4	1	[a]	[a]	[a]	[a]
4	5	1	[a]	[a]	[a]	[a]
4	6	3	2.83 (.29)	2.30 (1.12)	9.35 (5.62)	[a]

Note. Only scores on which raters agreed are included. SD in parentheses. $N = 113$.

[a] One or fewer scores at that level.

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