Research Report

Automated Scoring of Spontaneous Speech Using SpeechRater^{SM} v1.0

Xiaoming Xi, Derrick Higgins, Klaus Zechner,
and David M. Williamson

November 2008

ETS RR-08-62

Automated Scoring of Spontaneous Speech Using SpeechRater^{SM} v1.0

Xiaoming Xi, Derrick Higgins, Klaus Zechner, and David M. Williamson
ETS, Princeton, NJ

November 2008

Abstract

This report presents the results of a research and development effort for SpeechRater^{SM} Version 1.0 (v1.0), an automated scoring system for the spontaneous speech of English language learners used operationally in the Test of English as a Foreign Language™ (TOEFL®) Practice Online assessment (TPO). The report includes a summary of the validity considerations and analyses that drive both the development and the evaluation of the quality of automated scoring. These considerations include perspectives on the construct of interest, the context of use, and the empirical performance of the SpeechRater in relation to both the human scores and the intended use of the scores. The outcomes of this work have implications for short- and long-term goals for iterative improvements to SpeechRater scoring.

Key words: Automated scoring, automatic speech recognition, automatic speech processing, TOEFL, TOEFL Practice Online, and validity argument

Executive Summary

SpeechRater^{SM} Version 1.0 (v1.0) is an automated scoring system deployed for the Test of English as a Foreign Language™ (TOEFL®) Internet-based test (iBT) Speaking Practice Test, which is used by prospective test takers to prepare for the official TOEFL iBT test. This study reports the development and validation of the system for low-stakes practice purposes. The process we followed to build this system represented a principled approach to maximizing 2 essential qualities: substantively meaningful and technically sound. In developing and evaluating the features and the scoring models to predict human assigned scores, we engaged both content and technical experts actively to ensure the construct representation and technical soundness of the system. We compared primarily two alternative methodologies of building scoring models—multiple regression and classification trees—in terms of their construct representation and empirical performance in predicting human scores. Based on the evaluation results, we concluded that a multiple regression model with feature weights determined by content experts was superior to the other competing models evaluated.

We then used an argument-based approach to integrate and evaluate the existing evidence to support the use of SpeechRater^{SM} v1.0 in a low-stakes practice environment. The argument-based approach provided a mechanism for us to articulate the strengths and weaknesses in the validity argument for using SpeechRater v1.0 and put forward a transparent argument for using it for a low-stakes practice environment. In particular, the construct representation of the multiple regression model with expert weights was sufficiently broad to justify its use in a low-stakes application. While some higher-order aspects of the speaking construct (such as content and organization) are missing, more basic aspects of the construct (such as pronunciation and fluency) are richly represented. In addition, these different parts of the speaking construct tend to be highly correlated, so that the absence of higher order factors is not as detrimental to the model's agreement with human raters as it otherwise might be. The model's agreement with human raters was not sufficiently high to support high-stakes decisions but was still suitable for use in low-stakes applications. The correlation of the 6-item aggregate score with human raters was .57 and was deemed acceptable given the low human agreement and the fact that we obtained a much higher correlation of .68 on data with more variability in the scores, such as the iBT field study data. Furthermore, the dependability of the scores predicted by the final scoring model was quite high for the 6 tasks, supporting the high degree of generalizability of scores across tasks.

We also identified gaps in our existing research for SpeechRater v1.0. Specifically, the areas of research to pursue include improving the prediction accuracy for the whole test-taking population and for test takers with different native language backgrounds and expanding the construct coverage of the scoring model. Furthermore, we need to explore alternative criterion measures other than human scores to validate the scores provided by SpeechRater. Other critical areas of investigation include users' perceptions of and interactions with this system and the impact of users' perceptions on their decision making based on the scores.

Table of Contents

List of Tables

List of Figures

Page

1. Introduction

The multiple-choice item has flourished as the preferred item type for more than a generation, and with good reason. It is efficient to develop and administer, can be scored relatively unambiguously and swiftly, and is supported by a rich infrastructure of statistical methods and test theory. Include the increasing availability and lower cost of online administration and instantaneous scoring and it is no mystery why item types other than multiple choice are often not even considered as options for an assessment program.

Despite all these factors in favor of multiple choice, many areas of ability remain for which multiple-choice items alone are believed to result in an incomplete representation of the construct. For assessment of these constructs the choices traditionally have been to accept a swift and efficiently scored measure of a construct considered to be somewhat ill-fitting, using multiple-choice items alone, or to accept the slower and more costly human scoring that accompanies use of constructed-response items. Despite research into automated scoring of complex tasks extending back more than 40 years (e.g., Page, 1966), only relatively recently (Burstein et. al., 1998; Clauser, Margolis, Clyman, & Ross, 1997; Williamson, Bejar, & Hone, 1999) has the ability for computerized delivery and automated scoring of constructed-response items enabled the practical operational use of automated scoring for such items. Initially, such applications were primarily in automated scoring of essays (e.g., Attali & Burstein, 2006; Burstein et al., 1998; Chodorow & Burstein, 2004; Landauer & Dumais, 1997; Rudner, Garcia, & Welch, 2006), which has matured to a considerable degree. However, recent research in natural language processing and speech recognition capabilities has expanded the nature of constructed-response tasks that are automatically scorable to include short-answer tasks requiring factual information (e.g., Leacock & Chodorow, 2003) and tasks eliciting highly predictable speech (e.g., Bernstein, 1999).

This report presents the results of a research and development effort for SpeechRater^{SM} Version 1.0 (v1.0), an automated scoring system for the spontaneous speech of English language learners used operationally in the Test of English as a Foreign Language™ (TOEFL®) Practice Online assessment (TPO). This includes a summary of the validity considerations and analyses that drive both the development and the evaluation of the quality of automated scoring. These considerations include perspectives on the construct of interest, the context of use, and the empirical performance of the SpeechRater v1.0 automated scoring engine in relation to both the human scores and the intended use of the scores. The outcomes of this work have implications for short- and long-term goals for iterative improvements to SpeechRater scoring.

We start with the validation framework used for evaluating SpeechRater v1.0 (Section 2). We then describe the architecture of SpeechRater v1.0, its major components (Section 3). Following that, we describe the various data sets used in this study and discuss the development and validation efforts associated with each of the three components (Sections 4–8). In Section 9, we summarize and synthesize different lines of development and validation work presented in Sections 4–8 to support the validity of SpeechRater v1.0 and make a recommendation for using SpeechRater on a conditional basis. We conclude Section 9 by discussing the critical areas of future research that support enhancements of SpeechRater.

2. Validation Framework for SpeechRater v1.0

Validity Framework for Automated Scoring

Fundamentally, the validity considerations around the use of automated scoring in an assessment are no different from those for assessments that rely on human scoring of constructed-response items, or even those that use only multiple-choice items. The same obligations to ensure that the reported scores are appropriate for the assessment purpose persist, regardless of the mode of scoring or the type of item. However, automated scoring is often dependent on a complex set of technological mechanisms and algorithms for the production of a task score. This fact implies a greater responsibility to conduct a more thorough, critical evaluation of the quality of scoring than is typical (rightly or wrongly) for human scoring of similar tasks (a review of such validation efforts for automated scoring can be found in Yang, Buckendahl, Juszkiewicz, & Bhola, 2002). Therefore, we begin by integrating the considerations that are particular to automated scoring into a framework for validating an assessment as a whole, with an emphasis on the unique challenges and threats to validity of test scores specific to the application of automated scoring. What follows is a brief discussion of concepts and views of validation that have shaped our conceptualization and execution of the work conducted to validate the SpeechRater v1.0 scoring for use in the TPO assessment.

Validity and validation are fundamental notions in psychometrics and the subject of many papers, journal articles, and book chapters (e.g., Cronbach, 1971; Cureton, 1951; Kane, 2006; Messick, 1989). Validity is a theoretical notion that establishes the conceptual and empirical relationship between an assessment and its intended use. As such, it defines the expected meaning of a test score in relation to the construct or criterion the assessment is intended to represent and therefore the scope and nature of work required in order to establish the fundamental utility of assessments for their intended purpose. Validation is the process of developing and evaluating the evidence for and against the hypothesis that the assessment results are meaningful and appropriate for its intended use. Given the many perspectives and frameworks for validity and validation, it would be overly ambitious to present a full discussion of the considerations for automated scoring within each current or historical perspective. Instead, we adopt the position that the validation of automated scoring is, in general, an area of special consideration within the overall field of validity theory and practice in measurement. As a result, regardless of the overall validation framework adopted, there is a need for special consideration of the role of automated scoring within the chosen framework.

Most frameworks for validity emphasize (and rightly so) the validation of the reported scores from a test as the basis for supported decision making. As such, validation frameworks in general tend to focus on the basis for *inferences* (test scores) from assessment rather than on the fundamental elements of *evidence* (item scores) that contribute to the overall inference. Prior studies have tended to emphasize one or more of three approaches: (a) demonstrating the correspondence (in both agreement and reliability) in item-level scores between automated scoring systems (either a single system or multiple independent systems) and human scorers, (b) examining the relationship between item-level automated scores and criterion measures external to the assessment, and (c) understanding the construct represented within the scoring processes that automated scoring systems use (Yang et al., 2002). With few exceptions, these focus on item-level evaluations and restrict their scope to one of these three approaches to validation. Some have advocated the incorporation of multiple approaches into formal validity arguments for the overall assessment in a manner consistent with a more comprehensive validation strategy (e.g., Bennett & Bejar, 1998; Clauser, Kane, & Swanson, 2002). However, few, if any, have yet undertaken such comprehensive analyses combining both conceptual and empirical approaches to the extent advocated in these presentations.

In this evaluation of SpeechRater automated scoring for the TPO assessment, we adopt a conceptual validation framework presented by Clauser et al. (2002), within which the three areas of investigation mentioned above are integrated into a coherent evaluation of the appropriateness of scores produced by an automated scoring engine. This framework presents an argument-based approach to the validation of an assessment that uses automated scoring and is consistent with current argument-based presentations of validation (e.g., Kane, 1992, 2001, 2002, 2004; Kane, Crooks, & Cohen, 1999) that evolved from ideas presented by Cronbach (1988) and Messick (1989). Drawing on argumentation theories stemming from Toulmin logic (Toulmin, 2003), the current argument-based approach to validation has formalized the process of building and supporting arguments, thus offering a working framework for practitioners. Another application of Toulmin's argumentation theories is the evidence-centered design framework, which provides useful guidance for designing assessments in a principled way (Mislevy, Steinberg, Almond, & Lukas, 2006). The evidence-centered design approach provides a mechanism for building validity into an assessment from the outset, but a principled approach to assessment design is not sufficient to establish validity, which is best investigated within Kane's validation framework (Mislevy et al., 2006). In Kane and his associates' framework, validation is represented as a two-stage process consisting of the construction of an interpretive argument followed by the development and evaluation of a validity argument consistent with the interpretation.

Execution of a validation strategy within this framework begins with an articulation of the interpretive argument for each intended use of a test through a formal representation of the chain of inferences linking test performance to a decision, and the assumptions upon which these inferences rest. The stated assumptions, if supported through investigation, lend support for the pertinent inference. The inferences at the item level for such a representation are relatively modest. However, as the logical sequence progresses through the chain of inferences, each subsequent step results in inferences of greater and greater significance, and a correspondingly greater burden of evidence is required to support the inference, until the final inference of overall score or evaluation is represented. Theoretical, empirical, and procedural evidence of validity is used at each stage of the chain of inference to both question and support the assumptions that are required at that stage, with the emphasis shifting from conceptual and procedural at the more fine-grained levels (e.g., item level) to more empirical and rigorous at the level of the summary scores. The plausibility of the interpretive argument proposed for the assessment as a whole is based on this cumulative validity argument using theoretical and empirical evidence across the span of the chain of evidence. A strength of this approach lies in providing a transparent working framework to guide practitioners in three areas: (a) prioritizing different lines of evidence, (b) synthesizing them to evaluate the strength of a validity argument, and (c) gauging the progress of validation efforts. This approach also allows for a systematic way to consider potential threats to the assumptions and inferences and allocate resources to collect evidence intended to discount or reduce the impact of such threats.

Within this framework the validation of automated scoring is properly positioned as one specially targeted component of the overall validity argument for an assessment. Specifically, validation of automated scoring is a subset of the overall set of validation efforts. The evidence that automated scoring is an appropriate methodology at the item level and the test level supports the validity of the automated scores; this evidence must also be integrated into the overall argument for the validity of the assessment as a whole. As such, using this framework for validation positions, even demands, the work on automated scoring to be nested within considerations that extend beyond automated scoring to the other elements of an assessment design (e.g., Bennett & Bejar, 1998; Yang et al., 2002). Adopting the Clauser et al. (2002) framework provides an inherent representation of how decisions made in developing an automated scoring system may strengthen or weaken the overall validity argument, given the particular approach used to develop the system. This approach also allows for an explicit representation of potential threats to the strength of each inference in the chain that may be introduced by automated scoring and of lines of research that would alleviate or substantiate such threats. Although Clauser et al. (2002) did not cover all potential validity issues in validating a particular automated scoring system, they provided a useful working model for weaving the concerns directly associated with validation of automated scoring per se into this network of inferences, leading to a validity argument for the intended interpretation and use of test scores. Consequently, we have adopted this perspective in the current work. We will outline the role of this validation work on SpeechRater automated scoring within the overall validity argument for the TPO assessment in a subsequent section. However, before presenting the details of the validity study design for SpeechRater automated scoring for the TPO assessment, we first provide a brief summary of the assessment, including the desired claims and assessment structure, as the context for this research and the basis for positioning this effort within the overall validity framework for TPO per Clauser et al. (2002).

The TPO Assessment

The speaking section of the TOEFL Internet-based test (iBT) is designed to measure the academic English-speaking abilities of nonnative speakers who plan to study at English-medium institutions for higher education. Using tasks that require language use typical of an academic environment, the TOEFL iBT Speaking test represents an important advancement in the large-scale assessment of productive skills. However, it poses challenges to learners in parts of the world where opportunities to practice speaking English are limited.

The TPO assessment is designed to help prospective TOEFL iBT examinees become familiar with and better prepared for the test. As such, it is designed to mirror the content and design characteristics of the TOEFL iBT to the extent possible in an economical practice environment. As such, the various sections of the TPO each have the same number of items, same mixture of content represented, and same item types that appear in the operational TOEFL. In fact, all of the items that are used for the TPO are retired operational TOEFL iBT items. However, unlike the TOEFL iBT, the TPO allows users to customize their practice and take the test in a timed or untimed mode. The timed mode attempts to replicate the operational testing experience by using the same online delivery system and timing restrictions of the TOEFL iBT. In the untimed mode, users can progress at their own pace, starting or stopping the test whenever they like and revisiting items if desired. Another important distinction between the TPO and the TOEFL iBT is that the former targets more immediate and cost-effective score feedback; the TPO allows users to have immediate feedback on their performance for self-assessment of understanding of and comfort with the TOEFL iBT administration.

In early 2006 the users of TPO could instantly receive scores on reading and listening sections, both based on multiple-choice items, as well as the writing section, with automated writing scores provided by e-rater® (Attali & Burstein, 2006). The scores on speaking sections were produced by human raters within 5 business days. As a result of substantial interest in more immediate feedback from the speaking section of the TPO—the TOEFL iBT Speaking Practice test—a research agenda was launched to develop and deploy an automated scoring system for spontaneous speech. The immediate goal of this effort was to improve the scoring efficiency of the TOEFL iBT Speaking Practice test while maintaining quality comparable to that of trained human-rater scoring for the TPO assessment. The long-term goal is to provide instructional and diagnostic feedback based on automated features beyond the score feedback provided by human scoring while maintaining a level of validity of scores nearly equivalent to that of human scoring. The result of this effort was the release of SpeechRater v1.0 for operational use in the TPO.

The TOEFL iBT Speaking Practice test uses retired forms from the TOEFL iBT Speaking test. Like the TOEFL iBT, each test contains six tasks. The first two tasks are *independent* tasks that ask candidates to provide information or opinions on familiar topics based on their personal experience or background knowledge. The purpose of independent tasks is to measure the speaking ability of examinees independent of their ability to read or listen to English language. The remaining four tasks are *integrated* tasks that engage reading, listening, and speaking skills in combination to mimic the kinds of communication expected of students in campus-based situations and in academic courses. The entire test takes approximately 20 minutes. For each of the six tasks the examinee is allowed a short time to consider a response and then 45–60 seconds (depending on task type) to provide a spontaneous response.

The scoring rubric used by human raters to evaluate the responses to the TOEFL iBT Speaking Practice test is identical to that used for the TOEFL iBT Speaking test. Similarly, the scoring process uses human raters who rate the operational TOEFL iBT Speaking test. The raters issue a holistic score for each response on a score scale from 1–4 that is based on three key categories of performance: (a) Delivery, (b) Language Use, and (c) Topic Development. *Delivery* refers to the fluidity and clarity of the speech. In assessing Delivery, raters consider the speaker's pronunciation, intonation, rhythm, rate of speech, and degree of hesitancy. *Language Use* refers to the diversity, sophistication, and precision of vocabulary and the range, complexity, and accuracy of grammar. Raters evaluate candidates' ability to select words and phrases and their ability to produce structures that appropriately and effectively communicate their ideas. *Topic Development* refers to the coherence and fullness of the response. When assessing this dimension, raters take into account the progression of ideas; the degree of elaboration; the completeness; and, in the case of integrated tasks, the accuracy of the content. The rubric for human scoring represents the construct of speaking that is of interest to both the operational TOEFL iBT Speaking test and the TOEFL iBT Speaking Practice test.

One point worth mentioning is that the human scoring processes for the TOEFL iBT Speaking Practice test differ from operational scoring of the TOEFL iBT; the responses are scored task by task as they arrive, rather than in batches of responses that are all to a common task. In addition, the human raters are aware that they are scoring the practice test rather than the operational test.

The Validity Argument for Scoring TPO With SpeechRater v1.0

With the general validation approach established and the context of use defined, this section presents the specific validity argument to be supported for SpeechRater v1.0. As Clauser et al. (2002) noted, the decision to use automated scoring will impact not only the strength of the evaluation inference, which links test performance to observed test scores, but also the subsequent inferences in the validity argument. This is described as the “ripple effects” of automated scoring that “extend through each step in the argument” (Clauser et al., 2002, p. 420). To position automated scoring in a larger validity argument for the whole assessment, a general description of the chain of inferences resulting in assessment-based action (Kane et al., 1999, and Chapelle, Enright, & Jamieson, 2008) is provided below as Figure 1 illustrating the accumulation of strengths and weaknesses of the validity argument from the item level through the point of taking action on the basis of assessment results. Figure 1 also provides an illustration of the ripple effect of decisions and validity evidence at various stages of the chain of reasoning within a validity argument.

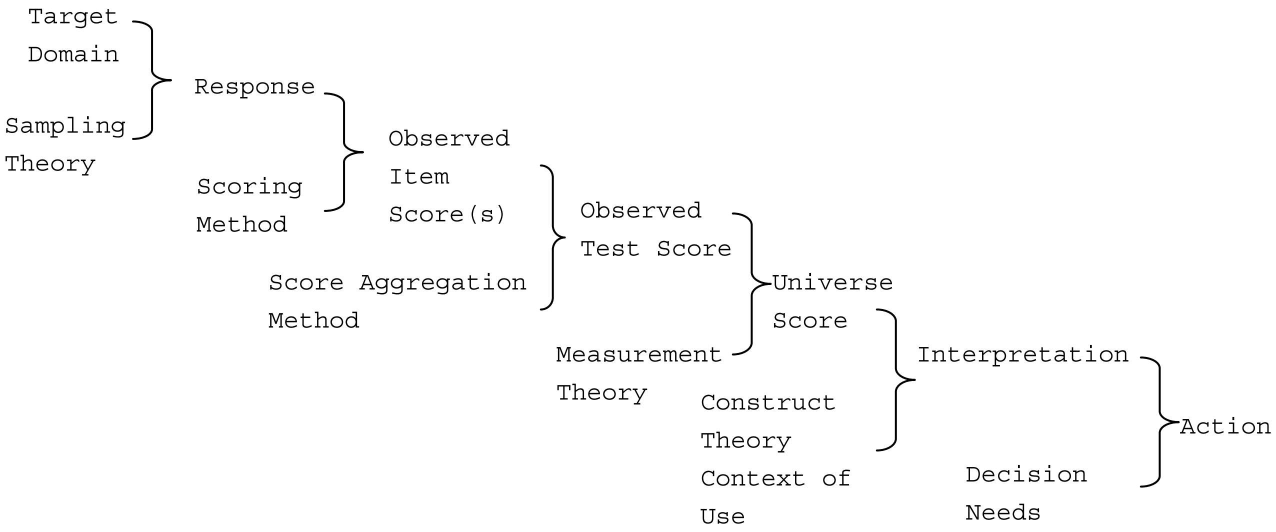


Figure 1. **Links in an interpretative validity argument.**

The validity argument starts with the most fundamental considerations in conceptualizing and developing an assessment, defining the target domain of interest and designing assessment tasks to elicit the knowledge, skills, and abilities that are intended to be measured. The target domain provides a basis for observations of performance on a test to reveal relevant knowledge, skills, and abilities.

The second stage of the validity argument involves the application of an item-level scoring methodology to a response to produce an observed item score. The scoring methodology may be straightforward, as in most multiple-choice item scoring methods, or it may be complex and require considerable effort, and evidence, to ensure that it is fully appropriate. Such complexity is the case in many scoring models that use human scoring to evaluate performance-based assessment tasks.

In the third stage the results of many such item-level scoring processes are combined through simple summation or averaging or by using more sophisticated measurement models, such as item-response models, to produce a total test score. This observed test score is typically the basis for most assessment-based decision making and the primary emphasis of validity studies. Exceptions include cases in which subscores are reported and are intended to be sufficiently reliable to take some kind of action on the basis of the results.

The fourth stage focuses on the consideration of the test score in combination with key aspects of measurement theory to recognize how well it would be expected to represent a universe score. Relevant interpretative concepts from measurement theory include true score estimation, reliability, and other principles that would be applied to estimate the degree to which the observed score is ultimately considered sufficiently accurate for decision making.

In the fifth stage, meaning can be attached to the universe score in two potential ways to support valid interpretations of the assessment results. The universe score can be interpreted by drawing on a theoretical construct (e.g., a communicative competence model) that underlies consistencies in test takers' performances. For assessments for which specific domains of generalization can be defined, this representation of the meaning of assessment results is further contextualized in the domain to which the test scores are intended to be generalized. The assessment results and inferred meaning, as influenced by both operational and theoretical constraints, are contrasted with the concerns and needs of the environment that the assessment simulates or predicts, which may be quite different for assessments used in certification and licensure than for those used for academic placement. This is the stage where domain and construct theories that dictate the test design blueprint, the principles of item and rubric development, and investigations of examinees' engagement with test items (processes and skills engaged) as moderated by the delivery mechanisms for assessment become highly relevant to the question of validation. Although these elements are not explicitly represented in Figure 1, they may constitute important evidence establishing the crucial link from a universe score to an interpretation.

Finally, the contextualized interpretation of observed test scores, viewed as a potential representation of a universe score, is applied in the context of a decision-making need to result in action based, at least in part, on the assessment results. Aspects of the decision need to include such considerations as targeted selection rates, consequences of misclassification, and other constraints and demands of the decision-making process. Actions include those for individuals who take the test as well as actions based on an understanding of large-scale assessment results that are intended to effect change for large groups of people or for policymaking.

As can be seen from this illustration, at the most global level the question of validity rests on whether the actions taken on the basis of assessment are warranted. Under this representation of a chain of reasoning for a validity argument, the implications of validity research required for an automated scoring system can be localized to the scoring method portion of the argument in the second stage of the validity chain. Specifically, the core issue of evaluation of automated scoring in the context of a validity argument for an assessment design is the replacement of a human scoring method with an automated one for the item-level scoring. The emphasis of validation efforts would be focused, initially, on the localized decision of item-level scoring but would be further obligated to evaluate the ripple effect, as such a replacement impacts the chain of reasoning downstream from the scoring method, with the most significant interest typically focused on implications for validity of reported scores for their intended use.

Figure 2 represents an overlay of the six types of validity arguments, discussed in the context of automated scoring by Clauser et al. (2002), onto the chain of validity represented by Figure 1. For each type of validity argument, an arrow indicates the part of the validity chain of reasoning that is the focal point of the argument.

The first type of argument, Domain Definition, is based on the premises that test tasks are sampled adequately from the target domain and that observations on a test are representative of the knowledge, skills, and abilities required in the target domain. Although this type of argument is central to the overall validity argument for an assessment, automated scoring does not typically come into play at this stage.

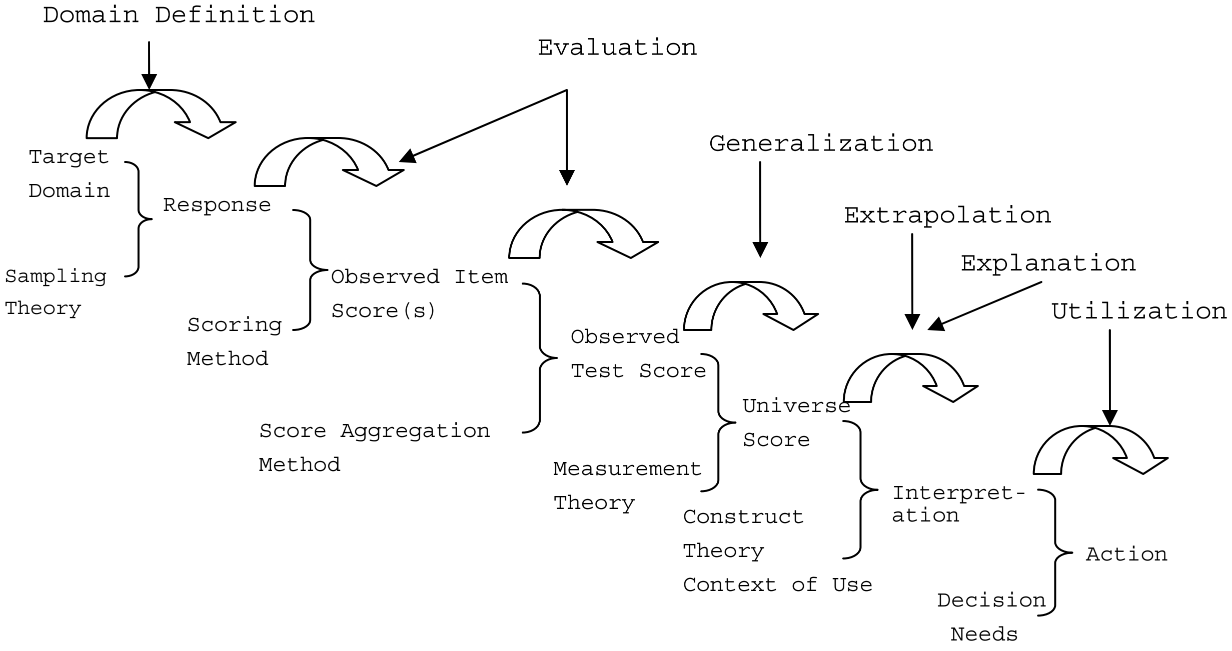


Figure 2. **Six types of validity argument and their focus.**

The second type of argument, Evaluation, emphasizes the acceptance of a score being conditional on the assurance that the score was assigned using appropriate procedures. While initially presented as a target for a test score, the extension of this concept to the item score is obvious, particularly for the question of automated scoring. Of course, the extension to the item level is an addition to help facilitate the overall Evaluation at the test-score level rather than to preclude it. In the context of the TPO assessment, the specific implication is of whether the examinee responses are obtained and scored appropriately so that the resulting score accurately represents the quality of the performance. For the specific question of use of automated scoring, there is a rich precedence of studies comparing the results of automated and human scoring as the basis for evaluating the validity of automated scores from the Evaluation perspective (e.g., Bejar, 1991; Bennett, Sebrechts, & Marc, 1994; Braun, Bennett, Frye, & Soloway, 1990; Clauser et al., 1995; Clauser et al., 1997; Kaplan & Bennett, 1994; Page & Petersen, 1995; Sebrechts, Bennett, & Rock, 1991; Williamson et al., 1999). It is generally acknowledged that human scores are not perfect and therefore may not be an ideal basis for evaluation of the quality of automated scoring, suggesting that the term *gold standard* may be a misnomer. However, human scores continue to be a highly relevant and convenient basis for evaluating the quality of automated scores. To support claims about the potential for improvements in quality of scoring resulting from automated scores, additional evaluation against criteria is needed, beyond direct comparison with a single set of human scores alone.

The third type of argument, Generalization, emphasizes the ability to generalize from the observed test score to a score that would be expected from some universe of observations. In this case, certain assumptions are made about the interchangeability of certain observations in the universe of possible observations that could conceivably be made, ultimately resulting in the assumption of invariance of the observed test scores resulting from different conditions of observation. The key goal of this area of emphasis is the extent to which the observed test score can be taken to be representative of a universe (true) score. The pertinent assumption for the specific case of language testing is that scores on language test tasks are generalizable over similar language tasks in the universe, raters, test forms, and occasions. In order to support this link, evidence is needed that the errors incurred in the measurement process are minimized to a level where we can be sure that, if a test taker were given similar language tasks, rated by different raters, or administered an alternate form or the same test on a different occasion, he or she would obtain a similar observed test score. The fundamental concerns and protocols for addressing this area of validity research are very similar between human and automated scoring.

The Extrapolation argument emphasizes the relationship between the tasks administered in an assessment, often artificial and contrived for test purposes, and the criterion tasks in a real-world environment in which the examinees are expected to function. This argument builds on the Generalization argument; however, here the emphasis has shifted to whether scores on test tasks could adequately predict performance in the real-world area of practice. This is crucial in the overall validity argument for a language assessment, because it bears on whether test takers' scores on the test provide adequate evidence about the language abilities that underlie their language performance in a target domain beyond the test. The assumptions are that test scores reflect the quality of language performance on relevant tasks in the real world. However, this is an initial assumption that demands an evidential basis through validation. For the specific concerns of automated scoring, a relevant question is whether there is any differential evidence of increased or decreased validity from the Extrapolation perspective as a result of substituting automated scoring for human scores.

Explanation is an area of validity research that focuses on the explanatory capacity of the assessment results. Specifically, this area of emphasis focuses not just on the score being an adequate predictor of performance level in the domain of interest, but also on the ability to draw explanatory inferences from the score beyond its ability to be predictive of criteria of interest. As a result, approaches to scoring that are construct based and capable of offering specific hypotheses about the abilities of examinees (such as diagnostic strengths or weaknesses, work processes, etc.) offer more contributions to the Explanation argument for validity than those that are less able to provide specific hypotheses beyond the association of scores with a limited range of criteria of interest. The Explanation area is relevant to automated scoring because it may be possible to obtain automated scores that are similar to those of human graders, but for reasons that are inconsistent with the meaning that the scores are intended to convey. As a case in point, consider that automated scoring of essays may produce scores that are fairly similar to those of human graders using nothing more than the essay length in words, yet the meaning imparted to the resultant score by such an approach would be unsatisfactory. A more construct-representative method of automated scoring would be preferred, even at the cost of some degree of association with human scores. As a result, the question of how scores are produced and the implications for the construct representation of such scores are a critical part of the overall validity evaluation of scores, both automated and human.

The final area of investigation, Action, is the direct connection between the score-based interpretations and the decisions made, at least in part, on the basis of assessment results. The assumptions are that the test scores and other related information provided to users are *relevant, useful,* and *sufficient* for making intended decisions and promote positive effects on teaching and learning (Bachman, 2005).

Some arguments for, and against, the use of automated scores have referenced these classes of validation efforts as the basis for expectations of strengths and weaknesses of automated scoring in operational settings. For example, with respect to the Evaluation aspects of validity, the use of automated scoring could make the logic of score production completely transparent and reproducible, offering the promise of standardization and openness to critique and modification as needed or feasible. Of course, with this advantage comes the concern that automated methods do not actually replicate the same cognitive processes that human graders undertake when they score responses, even when such processes are fairly well understood.

Therefore, the underlying logic of an automated system, while perhaps similar to human graders, remains different in key respects. The direct comparisons between human and automated scoring are the most typical and prevalent method for validation of automated scores, which is generally appropriate considering that such Evaluation is targeted at the point at which automated scoring has its most direct impact on score production: the scoring method.

Similarly, for Generalization the argument has been made that since an automated scoring system applies the defined rating criteria consistently, regardless of the circumstances of the solution, it can improve score generalizability by eliminating differences in the leniency or harshness of raters' judgments over tasks, occasions, or combinations of these. The potential threat to validity of such an advantage is that such consistency is only beneficial to the extent that there is confidence that the approach taken is appropriate for all possible responses. Any novel or unanticipated, but appropriate, solutions to task prompts may not be well handled by an automated system with a fixed and rigorous method for evaluating responses. Also, any errors or undesirable aspects of the automated scoring systems that do exist will be systematically applied to all responses, resulting in the potential for bias that also may compromise the Evaluation and Explanation aspects of the validation argument.

The area of Extrapolation is one in which use of automated scoring has relatively little impact beyond what is specified in the other areas of emphasis. A common argument in favor of automated scoring from the Extrapolation perspective is that the speed and efficiency of automated scoring can make the use of tasks with greater real-world fidelity more feasible for operational use. Counterarguments for use of automated scoring from the Extrapolation perspective include the concern that automated scoring is itself an artifact of the testing process and that in practice the work of examinees is not evaluated by an automated system but by other humans with whom they interact.

Explanation is an area in which both the strengths and weaknesses of automated scoring are generally apparent. A proposed strength is that automated scoring allows for the maximization of construct representation by selecting construct-relevant response features and combining them to produce scores in a way that best represents the construct (Bennett, 2006; Bennett & Bejar, 1998). This degree of control is not possible with human scoring. Similarly, the ability to have extensive and consistent analysis of a response promises to expand the potential for meaningful performance feedback with detailed descriptions of the strengths and weaknesses of the response that are not feasible under human scoring. By contrast, along with the potential advantage of customizing the construct representation of scoring with an automated scoring system, some aspects of the construct simply will not be captured to the satisfaction of experts in the field with automated mechanisms. This is particularly true for developing a scoring system for a construct as complex and challenging as speaking proficiency. Conceptualizing and implementing speech features that indicate the key criteria human raters use to score spoken responses present immense challenges. The tendency to extract easily quantifiable aspects of the performance due to the limitations of current speech technologies could result in construct-irrelevant features or features that do not represent the full construct of interest to the assessment. In addition, given the complexity of human raters' decision-making processes involved in rating speaking, designing a scoring system that adequately reflects those processes is obviously not an easy task. Even a scoring solution informed by expert judgments may not be adequate in representing the intended constructs, depending on the qualifications of the experts and the rigor with which the work is conducted.

Finally, the Utility area of validity work remains largely unchanged under automated scoring as it would be with human scoring. In this sense automated and human scores are subject to very similar concerns, both in terms of strength and weakness, as to how these pertain to the validity of assessment-based decision making. Despite the overall similarity of concerns, specific issues in Utility are relevant in the context of automated scoring: (a) whether the accuracy of the automated scores supported the intended decisions, (b) whether the knowledge of the scores being assigned by a computer will change the user's perception of the assessment and the way that user approaches the tasks and uses and interprets the results, and (c) whether automated scoring will promote positive effects on teaching and learning practices.

Since a validity argument is only as strong as its weakest link (Kane, 1992), it is critical to identify all the potential threats to the various inferences and provide counterevidence against the rebuttals. The validation efforts for SpeechRater v1.0 as a mechanism for scoring responses for the TPO focus on providing counterevidence that discounts these rebuttals. Therefore, to build and evaluate a validity argument for the SpeechRater v1.0, four basic steps are involved:

1. Clearly state the intended interpretation and use of the automated scores on the TOEFL iBT Speaking Practice test (the Action, from Figure 2).

2. Articulate the network of inferences that lead to the intended interpretation and use, consistent with the chain of reasoning outlined in Figure 2.
3. Identify critical rebuttals that may weaken each inference as a result of the use of automated scoring, based on the areas of validity research associated with the chain of reasoning in Figure 2.
4. Collect and integrate evidence to reject the potential rebuttals associated with each inference.

The first three steps will yield an interpretive argument, the plausibility of which will then be evaluated in Step 4 in the context of a validity argument.

The goal of SpeechRater v1.0 is to support the intended use of the TPO, to help students better prepare for and gauge their readiness to take the TOEFL iBT Speaking test. The claim at the observed item score level that is directly supported by use of SpeechRater v1.0 is: The SpeechRater v1.0 item score is a prediction of the score on the TOEFL iBT Speaking Practice test this response would have obtained from trained human raters.

The claims that are supported at the Utilization level for the speaking portion of the TPO, with scoring provided by SpeechRater v1.0, are the following: The TPO Speaking score, using SpeechRater v1.0, is a prediction of the score on the TOEFL iBT Speaking Practice test this examinee would have obtained from trained human raters. The entire practice experience can help familiarize test takers with the content and format of the TOEFL iBT Speaking test so that they can better prepare for it. This score can be used by the test takers to help them self-evaluate their readiness to take the TOEFL iBT Speaking test.

Table 1 shows the most common types of inferences that need to be verified to support the claims we would like to make based on scores generated by the SpeechRater v1.0. These are classified by general areas of validity research referenced above. The crucial rebuttals that may undermine the validity of the SpeechRater v1.0 are also stated, associated with the most pertinent validity area. Failure to provide evidence to reject any of these rebuttals could weaken the argument for the use of automated scoring. Within this framework this paper presents results of investigation of four major areas of emphasis:

1. *Evaluation*. The extent to which the scores provided by SpeechRater can be argued to be a reasonable prediction of human scores. The primary evidence for this claim is

Table 1

Areas of Emphasis for Validity of SpeechRater v1.0 and Associated Rebuttals

Inferences	Rebuttals
Evaluation: Automated scoring results in scores that accurately represent the quality of the performance on the practice test.	1. The scoring algorithm under- or misrepresents the construct or introduces construct irrelevance so that the resulting scores are not accurate.
Generalization: The scoring model can generalize to new tasks and samples of candidates, and the automated scores are generalizable over tasks.	1. The scoring model is built from insufficient or unrepresentative samples. 2. The scoring model does not generalize to new tasks or independent candidate samples. 3. The automated scores do not generalize across tasks.
Extrapolation: The automated scores reflect the quality of performance on relevant real-world speaking tasks in an academic environment.	1. Candidates' automated scores are not related to their levels of performance on real-world speaking tasks in an academic environment.
Explanation: The automated scoring model captures aspects of performance that reflect the underlying speaking abilities used in an academic setting.	1. The automated scores are not adequate in *explaining* examinee performance in the domain. 2. The speech features used in scoring models are not well linked to the rubric, introducing construct irrelevance. 3. The speech features do not cover the key criteria defined in the rubric very well, resulting in construct underrepresentation. 4. The speech features are not combined in a meaningful way to produce scores. 5. The scoring model disproportionately captures aspects of the rubric that generalize across tasks, reducing task specificity in an undesirable way, so that the constructs are underrepresented.
Utilization: The automated test scores and other related information provided to candidates are relevant, useful, and sufficient for them to make intended decisions and promote positive effects on teaching and learning.	1. The predicted scores and other information communicated to the candidates do not provide relevant, useful and sufficient information for them to gauge their readiness to take the TOEFL iBT Speaking test. 2. The automated scores negatively impact users' perceptions of the assessment and the way they interpret and use the scores as intended. 3. The automated scoring system does not promote positive washback effects on English language teaching and learning. 4. The potential negative consequences of SpeechRater v1.0 are not anticipated and minimized.

based on various empirical measures of association between automated and human scores on a common data set.

2. *Generalization*. Arguments for the generalizabilty of the SpeechRater v1.0 automated scoring to a variety of task prompts is provided on a conceptual basis, with respect to design decision in the construction and application of the scoring models as well as on an empirical basis with respect to the generalizability of the automated scores across tasks, estimated using generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972).
3. *Explanation*. Arguments for the appropriateness of SpeechRater v1.0 for explanation are provided and counterarguments presented on a predominantly logical basis, using the design decisions, features used in model construction, and the way features are combined to produce automated scores as the basis for such arguments.
4. *Utilization.* Arguments for the usefulness of the SpeechRater v1.0 scores for self-evaluations of readiness to take the official test are supported by an analysis of the magnitude of the prediction error in relation to the intended decision. Arguments about potential consequences of the SpeechRater v1.0 are made based on the score report and on the advisory information communicated to the user about the limitations of the system and the intended use of the scores, included as part of the user interface.

Other evaluations that are clearly relevant and necessary for an overall validity argument, including the evaluation of relationships of automated scores with external measures of ability, are targeted for future work. To give the reader a full understanding of exactly what is being validated in this study, the next section provides a schematic of the organization and operation of the SpeechRater v1.0 speech scoring system.

3. Architecture of an Automated Scoring System

This section describes the architecture of an automated speech scoring system, which serves as a natural organizing structure for the remaining of the paper. An automated speech scoring system consists of three major components (see Figure 3). The speech recognizer and the feature generation programs are closely interrelated and can be considered as one big component that generates the scoring features. The speech recognizer decodes the input audio files into recognized words and utterances; then, the feature generation programs extract the scoring features indicating different aspects of performance in a response, based on various output that the speech recognizer produces. The second component is the scoring model used to score responses to individual tasks based on the scoring features and to summarize the scores across multiple tasks. The last component is the user interface that provides the score report and advisory information to users.

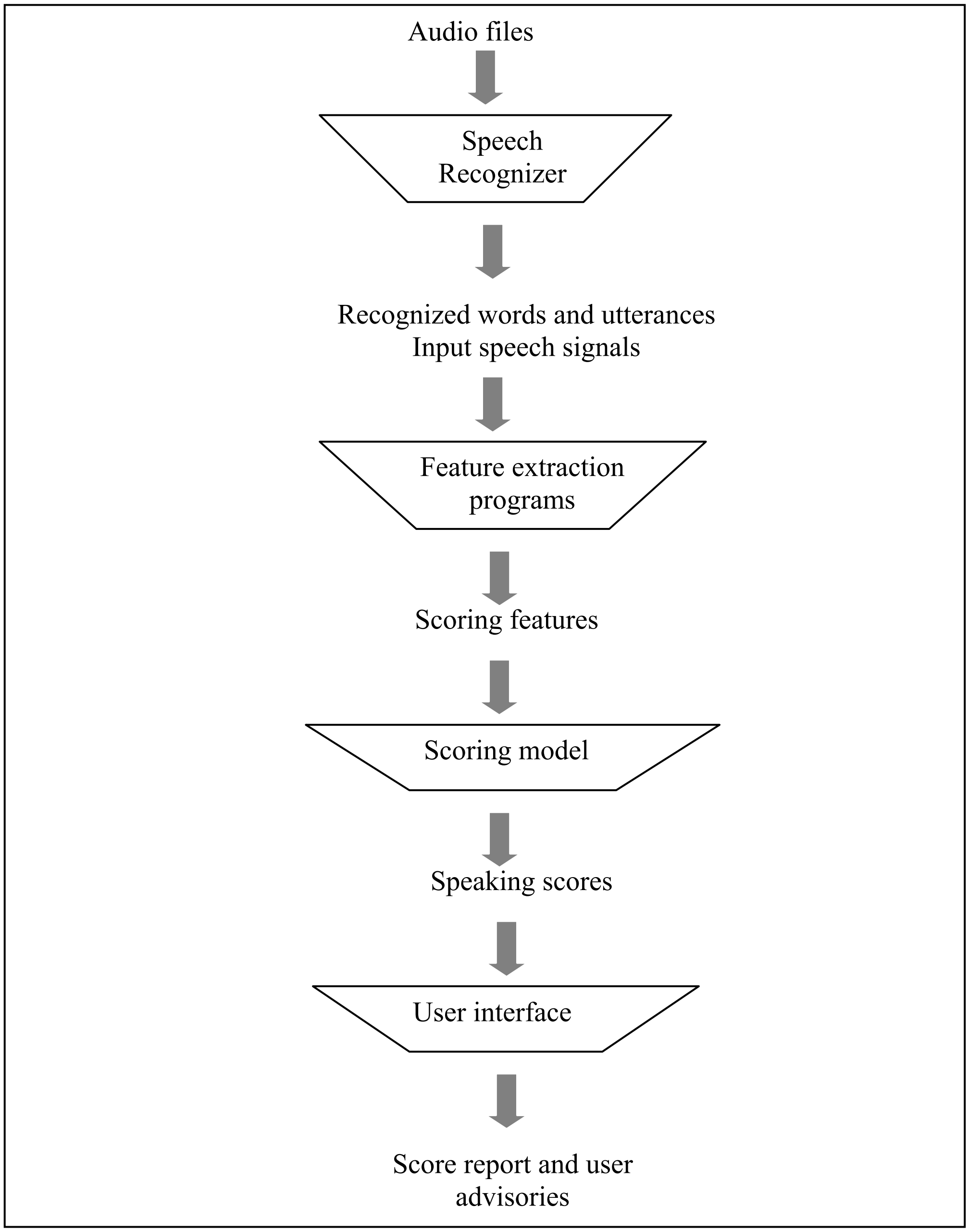


***Figure 3.* Architecture of an automated speech scoring system.**

In the next sections, we first describe the various data sets that were used to evaluate the speech features under consideration and to develop and evaluate the scoring model. We then address in sequence the development and validation efforts associated with each of the three major components of SpeechRater: (a) the scoring features, (b) the scoring model, and (c) the user interface. These three areas of investigation pertain to different inferences in the argument, with the first area providing evidence for the Explanation inference; the second one the Explanation, the Evaluation, and the Generalization inferences; and the third one the Utilization inference. In Section 9, the different lines of evidence are organized and synthesized to evaluate the soundness of the validity argument. On the basis of these evaluations a summary recommendation is described for using the SpeechRater in the TPO assessment and the implications of that recommendation summarized. The paper closes with recommendations for further work and extended validation studies.

4. Data

In building and evaluating the scoring models described in this report, we made use of two data sets: responses to the TPO assessment (the *TPO data set*) and responses from the TOEFL iBT Field Study (the *iBT data set*).

TPO Data

In total, the TPO data contained 4,162 spoken responses. An additional human score was obtained on each response as part of a special, intensive, human-scoring job. Nonadjacent discrepancies were adjudicated by a scoring leader in the special rating effort. For the purposes of model building and analysis, we used the second set of human scores, because they were undertaken under more optimal rating conditions. The adjudications were not used in the process of model building but were undertaken only to match our operational procedures for constructed-response scoring. The additional ratings might also be useful in future analyses.

The TPO data contained responses from four distinct test forms, with each test form containing six distinct speaking prompts: two independent tasks and four integrated tasks (see Section 3). Each TPO response may be assigned a score in the range of 1–4, or 0 if the candidate makes no attempt to answer or produces a few words totally unrelated to the topic. Each response also may be labeled as “technical difficulty” (TD) when technical issues may have degraded the audio quality so that a fair evaluation is not possible. These scoring rules are in accordance with the scoring of the operational TOEFL iBT, and with the scoring of the iBT field study data described below.

We set aside a portion of the TPO data for the training of the speech recognizer (the *rec-train* set, about 1,900 responses). The remaining data were partitioned into the scoring-model training (*sm-train*, about 1,300 responses) and scoring-model evaluation (*sm-eval,* about 500 responses) sets to maximize utility in evaluating the features and in building and evaluating the scoring models discussed in the next section. (The remaining responses were TD or 0 and were treated separately; see Section 9.) The sm-train and sm-eval sets consist of a set of responses with human scores in the range 1–4. The sm-train data were also used in evaluating the statistical properties of features (see Section 7) so that the feature selection was not biased by using the sm-eval data.

The partitioning of the TPO data was done in such a way that no overlap between speakers or tasks was allowed between the sm-train and sm-eval sets (to prevent overtraining on construct-irrelevant aspects of the response). The partitioning was also designed to minimize speaker and prompt overlap between the rec-train set and all other sets, although this constraint could not be enforced absolutely. In order to ensure that all data partitions were of sufficient size for their intended purposes, while meeting our other constraints, we were forced to accept some speaker and task overlap between the rec-train partition and other partitions. The total proportion of responses with task and speaker overlap with the rec-train set amounted to 25% of the sm-train set, and 31% of the sm-eval set. Because there was still no overlap between the sm-train and sm-eval sets, it is unlikely that this would result in inflated estimates of scoring accuracy for SpeechRater. There is a danger that the overlap with the rec-train set could artificially inflate the recognizer’s word accuracy on each of these other partitions. However, this is unlikely to have a large effect on the scores produced by the model, because our set of scoring features do not depend strongly on the accurate recovery of the words spoken in a response.

The partitioning process was also designed to ensure that the sm-train and sm-eval sets contain (a) a broad set of prompts, (b) similar proportions of responses from speakers of particular linguistic backgrounds, and (c) approximately the same proportion of responses to independent and integrated topics. This resulted in the division of the TPO data scored in the range of 1–4 into three sets, as shown in Table 2.

Table 2

Summary Statistics of TOEFL Practice Online Data Scored in the Range of 1–4

Data set	No. responses	No. speakers	No. topics	Average score	*SD* of score	Score distribution			
						1	2	3	4
rec-train	1,907	320	24	2.81	0.72	52 (2.5%)	550 (28.8%)	1011 (53.0%)	294 (15.4%)
sm-train	1,257	263	15	2.74	0.77	58 (4.6%)	405 (32.2%)	603 (48.0%)	191 (15.2%)
sm-eval	520	120	9	2.73	0.69	18 (3.5%)	159 (30.6%)	289 (55.6%)	54 (10.4%)

Note. Rec-train = speech-recognizer training, sm-train = scoring-model training, sm-eval = scoring-model evaluation.

The agreement between human raters on rating scorable responses in the 1–4 range was fairly low. Exact agreement was only 57.2%, with a quadratic-weighted κ of .554 and Pearson *r* of .55. The level of human agreement improved somewhat as we aggregated scores; the agreement on summed pairs of scores, triples, and full sets of six is presented in Table 3. As mentioned above, because the second set of raters rated under more optimal conditions, and the bulk of response adjudications tended to agree with them, we decided to use the second human ratings in doing our model development and evaluation.

Table 3

Human Agreement on Aggregated Scores for the TOEFL Practice Online Scoring-Model Evaluation and Speech-Recognizer Training Sets

No. of scores	Exact agreement	Exact + adjacent agreement	Quadratic-weighted kappa	Pearson *r*
1	57.2%	97.5%	.54	.55
2	40.0%	81.4%	.61	.63
3	28.8%	69.8%	.62	.68
6	15.5%	48.5%	.71	.74

Note. Technical difficulty and 0 scores omitted.

TOEFL iBT Field Study Data

The TOEFL iBT Field Study was a pilot study undertaken before the official roll-out of the TOEFL iBT. While we were primarily interested in model performance on TPO data, we used the field study data in doing some evaluation runs for a number of reasons. First, the conditions under which the field study data were scored were closer to best practice than they were to the TPO data sets. Additionally, the partitioning of the field study data allowed for better evaluation of the effects of item score aggregation, since the evaluation set contains more complete forms (sets of six tasks for a given examinee). Finally, evaluation of the field study data provided us with some idea of how our model generalizes across populations and audio file formats.

The field study data contained 3,502 responses from a single TOEFL iBT Speaking test form that were scored 1–4 (0s and TDs were not included). Since we did not need to train a new recognizer for these data, all of the data were used for the sm-train and sm-eval sets. These two sets of data were constructed to maximize the number of examinees with six complete tasks in a set so that we could evaluate candidates' total scores on this section. This constraint prevented us from enforcing a ban on task overlap between the sm-train and sm-eval sets but did allow us to prevent speaker overlap. Table 4 shows the properties of these two data sets.

Not all of the responses in these sets were double-scored, so we were forced to evaluate the level of human agreement on that subset of the data that had been double-scored. These results are provided in the Table 5. (Note that we did not have enough double-scored responses to provide agreement results for sets of six tasks.)

Table 4

Summary Statistics of TOEFL Internet-Based Test Field Study Data Sets

	No.	No.	No.	Avg.	*SD* of	Score distribution					
Data set	responses	speakers	topics	score	score	TD	0	1	2	3	4
sm-train	1,750	311	6	2.44	1.02	0	0	366	573	482	329
sm-eval	1,752	315	6	2.48	1.00	0	0	339	553	542	318

Note. TD = technical difficulty, sm-train = scoring-model training, sm-eval = scoring-model evaluation.

Table 5

Human Agreement on Aggregated Field Study Scores

No. of scores	Exact agreement	Exact + adjacent agreement	Quadratic-weighted kappa	Pearson *r*
1	57.1%	98.3%	.77	.77
2	45.1%	87.6%	.86	.86
3	36.7%	86.7%	.93	.94

One point to note is that the human–human agreement, as indicated by the weighted kappa and the correlation, was much higher for the field study data than for the TPO data. This reflects in part the fact that the field study scores were more varied and more evenly distributed across the four score levels than the TPO scores. In contrast, in the TPO data, the scores clustered around 3, with very few at the score level of 1. After adjusting the marginal totals of the TPO sm-eval human–human score matrix to mimic the distribution of marginal totals similar to that in the field study data (Haberman, 1979), the weighted kappa estimates increased from .55 to .76, and correlations between the two human ratings increased from .56 to .76 on single tasks for the TPO data.

The TOEFL iBT Field Study data are in a different file format than that of the TPO data. For this reason, all experiments using this data were run with a different recognizer, which was trained on candidate responses to the TOEFL Academic Speaking Test, a stand-alone test identical in content to the TOEFL iBT Speaking test and made available to prospective test takers for practice purposes. This recognizer is substantially similar to the TPO recognizer, as it relies on the same core software base, but was optimized for a different population of speakers and file format.

5. Development and Validation of Scoring Features

This section discusses the development, evaluation, and selection of the features used in the scoring models for SpeechRater v1.0 for TPO, focusing on the processes and strategies we employed to ensure the construct relevance, construct coverage, and empirical value of the features. The results presented in this section lend support to the Explanation area of inquiry, which attaches meaning to the SpeechRater scores by explicating the relationships between the scoring features and the speaking construct. Before the scoring features are discussed, some general background information on speech recognition and scoring systems and a description of a typical speech recognizer are provided.

Background on Speech Recognition and Scoring Systems

The technologies that support the automated evaluation of speaking proficiency are automated speech recognition (ASR) and analysis technologies as well as natural language processing tools (for a survey of these technologies, see Jurafsky & Martin, 2000). The application of these technologies to the TOEFL iBT Speaking Practice test poses challenges because this test elicits extended, spontaneous speech rather than highly predictable speech, and the scoring rubrics on which the responses are evaluated draw on models of communicative competence rather than adherence to an expected pattern of pronunciation, vocabulary, and fluency for highly predictable speech. In addition, the TOEFL iBT Speaking test has been developed to support the learning and teaching of academic speaking skills. Therefore, an automated evaluation system would benefit from having the potential to provide feedback on test takers' performances. To meet those challenges, it is necessary to ensure the accurate recognition of spontaneous, accented speech and the use of meaningful construct-relevant features extracted from the responses that represent various aspects of the rubric for score prediction.

The application of speech recognition and processing technologies to automated evaluation of speech is a fairly recent development. One successful area of research has focused on the automated evaluation of the pronunciation of nonnative speakers. Franco et al. (2000) have developed a system, EduSpeak, for the automatic evaluation of pronunciation by native and nonnative speakers of English and other languages at the phone and sentence levels. Candidates read English texts, and a forced alignment between the speech signal and the ideal path through the hidden Markov model (HMM) is computed. Based on this, the log posterior probabilities for pronouncing a certain phone at a certain position in the signal are computed to yield a local pronunciation score. This score is then combined with other automatically derived measures, such as the rate of speech and the duration of phonemes, for an overall pronunciation evaluation, using human judgments as the criterion.

A company called Ordinate (now part of Pearson Education, Inc.) has developed an automated scoring system based on highly constrained speech, elicited through test tasks such as reading sentences aloud, repeating sentences, or answering questions that require short responses containing only a few words (Bernstein, 1999). Scores in sentence mastery, fluency, pronunciation, and vocabulary based on these tasks are provided by means of speech-recognition and processing technologies. The Ordinate speaking assessments differ from the TOEFL iBT Speaking test in important ways. First, the TOEFL iBT test has been designed to support the teaching and learning of academic English. Therefore, the tasks used in the TOEFL iBT Speaking test elicit extended spontaneous speech typical of that used in academic courses and campus life. By contrast, the speaking tests developed by Ordinate aim to *predict* speaking proficiency. The tasks used in their assessments, constrained by the limitations of current speech-recognition and processing technologies, do not call for extended, spontaneous speech production and thus do not *directly* measure communicative competence. Second, for alignment with the goal of the TOEFL iBT test to promote good teaching and learning, the long-term goal of developing an automated scoring system for the practice test is to provide feedback on students' performances on the TOEFL Speaking Practice test that is useful for them to improve their speaking skills. This requires that their speaking performances be characterized by combining in a reasonable way meaningful speech features that represent various aspects of the rubric.

There is a growing body of research in the automated analysis of spontaneous speech, rather than constrained speech, which informs the effort to meet these goals. Cucchiarini, Strik, and Boves (2000) and Strik and Cucchiarini (1999), for example, have focused on the fluency features of free speech that can be extracted automatically from the output of a typical ASR engine. Their work has been influential in the conceptualization and implementation of relevant fluency features for this effort. However, the current effort targets more than fluency and aims instead to represent a full speaking construct suggested by current models of communicative competence (Bachman, 1990; Bachman & Palmer, 1996) and used as the model for scoring TOEFL tasks. At ETS, Zechner, Bejar, and Hemat (2007) made an initial effort to use speech technologies to extract speech features in fluency, vocabulary, and content that provided some evidence about the overall quality of responses to TOEFL iBT prototype speaking tasks, as indicated by the human scores. Xi, Zechner, and Bejar (2006) extended this effort by using an expanded and modified set of speech features that are more aligned with the human scoring rubric to predict both human holistic scores and analytic scores (Delivery, Language Use, and Topic Development) on TOEFL speaking tasks. This prior work formed the basis for the present study.

Description of the Speech Recognizer and its Role in Providing Data for Feature Extraction

An ASR system can be conceptualized as a system whose input is a digitized acoustic signal and whose output is the best estimate as to what sequence of words corresponds to the input signal (Jurafsky & Martin, 2000; Rabiner, 1989; and Rabiner & Juang, 1993, offer introductions to speech recognition). In an automated speech scoring system, the speech recognizer provides data describing some properties of the speech (e.g., recognized word strings, timing information associated with each word and between words, etc.). Based on these data, speech features such as fluency or vocabulary can be computed.

The ASR systems for the consumer market are commonly used for dictation and so are typically designed to maximize transcription accuracy. Therefore, an important metric in comparing ASR systems is their *word error rate,* which is based on a string alignment between the correct word string and the recognized word string (Jurafsky & Martin, 2000).

The architecture of an ASR system has become fairly standardized. The goal of such a system can be seen as transcribing the acoustic signal into a textual representation and is mediated by two models, the acoustic model (AM) and the language model (LM), as well as by a pronunciation dictionary. The AM associates probabilities with speech units called *phones* that represent a given phoneme. Phonemes are idealized representations for actually occurring sounds (phones) that fall—in articulatory and perceptual respects—into the same class of sounds. For example, one such phoneme may be realizations of an "ih" sound in English (e.g., in a word like *tip*).

The individual phonemes as well as sequences thereof are modeled with hidden Markov models, which can be understood as networks of nodes connected with directed and labeled arcs (*transitions* and transition probabilities); the nodes emit certain observations with associated observation probabilities. In our case, the observations are feature vectors derived from the digitized input signal. The task of a search algorithm is then to recover the underlying, hidden sequence (of phonemes and their parts) given a particular sequence of feature vector observations.

The second model of a speech recognizer is the LM, which models the prior probabilities of word sequences in English that are called *n-grams*. For example, a *trigram* is a sequence of three words where the probability of the third word occurring in the context of the first and second word is estimated.

Both the AM and LM models need to be estimated (trained) ahead of time. For the AM training, a reasonably sized, transcribed corpus of speech is needed that is very similar in accent and acoustic conditions to the speech to be expected in the operational system. For the LM training, similarly, a reasonably sized text corpus is needed that corresponds well to word sequences expected to be encountered in the operational ASR system, ideally in style, grammar, and vocabulary.

Finally, a pronunciation dictionary needs to be built where every word of the chosen vocabulary of the recognizer needs to have at least one associated pronunciation in terms of a sequence of phonemes. Some words may have a number of alternative pronunciations. For example, *little* could be phonetically transcribed as /l I t l/ or as /l I l/, where *l*, *I*, and *t* stand for the phonemes for the sounds of L, I, and T.

The AM, LM, and pronunciation dictionary are used jointly to decode the signal. The decoding process is a search through alternative transcriptions of the signal in order to locate the most likely transcription. The search mechanism is computationally complex, as the beginnings and ends of words are not given in advance. Therefore, different word boundaries have to be considered to determine a ranked list of possible transcriptions.

The performance of an ASR system is affected by a variety of factors related to generalizability, that is, the degree to which the training data are representative of the speech to be recognized. For example, the environment where the speech sample is captured, the surrounding noise level, the type and quality of the microphone, and the resolution of the speech signal are reflected in the AM. If those conditions are not maintained while using the ASR system after training, the performance of the ASR system will degrade when recognizing new speech. Other acoustic factors impact the performance of the AM, including the degree of accentedness and other speech idiosyncracies. The LM also affects the performance of the ASR system. The accuracy of the probabilities of observing n-grams clearly depends on the amount of the training data and on the match between the training and testing data in content and other characteristics.

The Construct of Interest That Motivates the Scoring Features

The TOEFL iBT Speaking test measures test takers' ability to speak about everyday familiar topics; to summarize, synthesize, and integrate written and audio materials; and to present the information orally in a comprehensible, coherent, and appropriate manner. The scoring rubric for human grading represents the construct of speaking that is of interest to both the operational TOEFL iBT Speaking and the TOEFL iBT Speaking Practice test. The full elaboration of this construct of speaking is provided graphically as Figure 4.

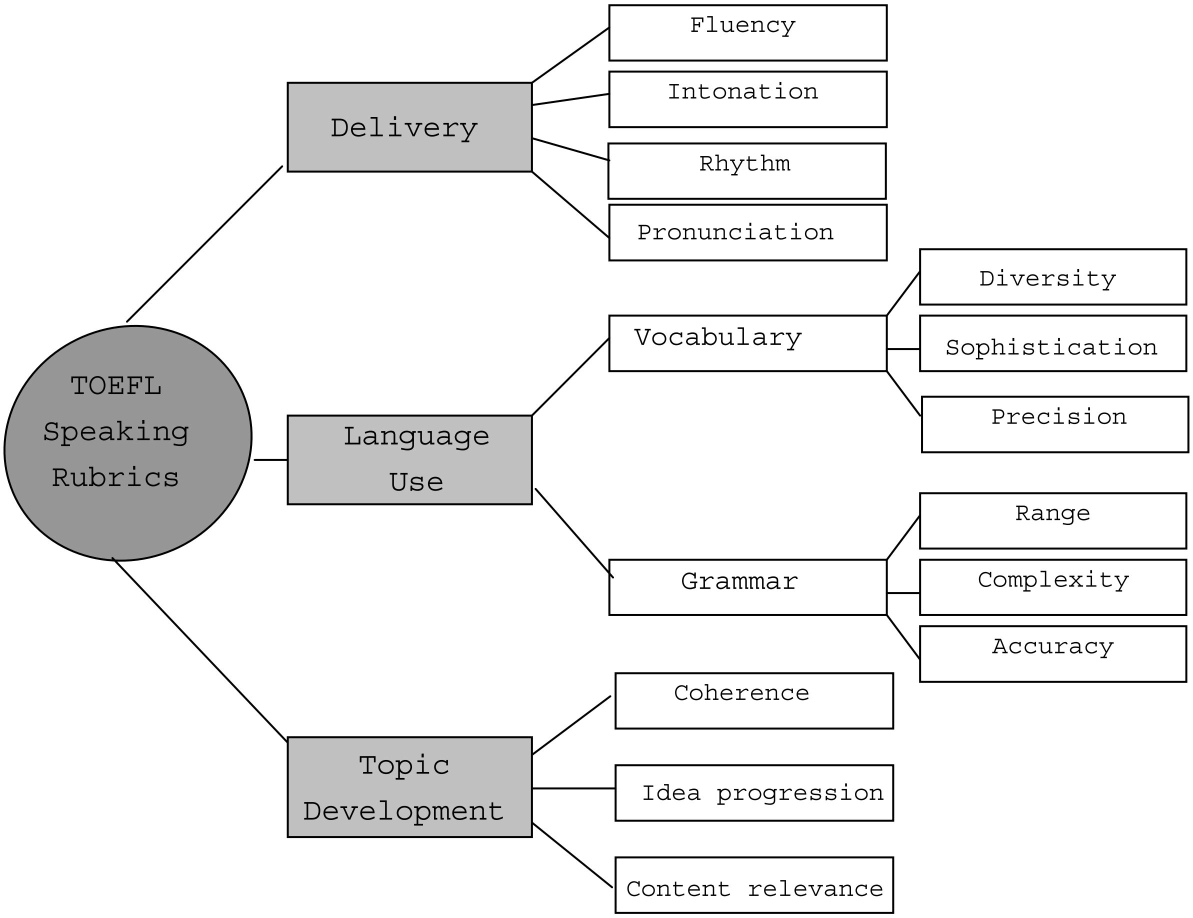


***Figure 4.* The construct of speech for the TOEFL Internet-based test represented by the scoring rubric.**

Delivery refers to the pace and clarity of the speech. In assessing Delivery, raters consider the speaker's pronunciation, intonation, rhythm, rate of speech, and degree of hesitancy. *Language Use* refers to the diversity, sophistication, and precision of vocabulary and the range, complexity and accuracy of grammar. Raters evaluate candidates' ability to select words and phrases and their ability to produce structures that appropriately and effectively communicate their ideas. *Topic Development* refers to the coherence and fullness of the response. When assessing this dimension, raters take into account the progression of ideas; the degree of elaboration; the completeness; and, in the case of integrated tasks, the accuracy of the content. Based on Brown et al. (2005), the rubrics for the TOEFL iBT Speaking test were reflective of what teachers of English as a second language and applied linguists thought were important in evaluating candidates' speaking performance in an academic environment. The construct of interest (and basis for scoring spoken responses) for the TOEFL iBT Speaking Practice test represents the basis for evaluating the degree to which the automated scoring of SpeechRater v1.0 is consistent with or deviates from representation of this construct.

Outline of the General Process Used to Derive Features

The candidate features were derived based on the rubric for the TOEFL iBT Speaking test and informed by the relevant literature and extensive feedback from assessment specialists and expert raters. We started with a detailed analysis of the rubric, with the goals of decomposing it into the feature classes shown in Table 6 and formulating features for each class that could be realized computationally by means of speech and natural language processing technologies. We then conducted an extensive literature review of the relevant literature in second language learning and computational linguistics. We focused on the body of literature on the linguistic analysis of speech in terms of fluency, accuracy, and complexity. Many of the variables used in the linguistic analysis of speech are manually coded, but they do provide a conceptual basis for us to evaluate them for our purpose and to formulate ways to compute them.

The computational linguistics literature yields some candidate features for us to evaluate and choose. Based on the literature review, we created a list of potential features indicating each feature class. We then obtained feedback from a group of content specialists about the meaningfulness of the features. Based on their input, we implemented the features that were considered well linked to the rubric.

Inventory of the Features and Their Linkage to the Construct

The Features

A total of 29 features were computed based on the outputs of the speech recognizer (Table 6). They are the candidate features that were evaluated and selected to develop the scoring models.

Table 6

Candidate Features for the Development of the Scoring Models

Feature	Feature class	Dimension	Description
1. Numwds	Length		# of words
2. Numtok	Length		# of tokens [numwds + numdff]; disfluencies counted as tokens
3. Globsegdur	Length		Duration of entire transcribed segment, including all pauses
4. Segdur	Length		Total duration of segment without disfluencies & pauses
5. Uttsegdur	Length		Duration of entire transcribed segment but without interutterance pauses
6. Wdpchk	Fluency	Delivery	Average length of speech chunks
7. Secpchk	Fluency	Delivery	Average duration of speech chunks
8. Wpsec	Fluency	Delivery	Speech articulation rate
9. Wpsecutt	Fluency	Delivery	Speaking rate
10. Secpchkmeandev	Fluency	Delivery	Mean absolute deviation of speech chunks in seconds
11. Wdpchkmeandev	Fluency	Delivery	Mean absolute deviation of speech chunks in words
12. Numsil	Fluency	Delivery	# of silence events
13. Silpwd	Fluency	Delivery	Duration of silences normalized by response length in words
14. Silpsec	Fluency	Delivery	Duration of silences normalized by total word duration
15. Silmean	Fluency	Delivery	Average duration of silences
16. Silmeandev	Fluency	Delivery	Mean deviation of silences
17. Longpfreq	Fluency	Delivery	Frequency of long pauses
18. Longpmn	Fluency	Delivery	Mean duration of long pauses

(Table continues)

Table 6 (continued)

Feature	Feature class	Dimension	Description
19. Longpwd	Fluency	Delivery	Frequency of long pauses normalized by response length in words
20. Longpmeandev	Fluency	Delivery	Mean deviation of long pauses
21. Silstddev	Fluency	Delivery	Standard deviation of silence durations
22. Longpstddev	Fluency	Delivery	Standard deviation of long pauses
23. Numdff	Fluency	Delivery	# of disfluencies (“uh,” “um”)
24. Dpsec	Fluency	Delivery	Disfluencies per second
25. Repfreq	Fluency	Delivery	# of repetitions normalized by response length in words
26. Tpsec	Fluency & vocab. diversity	Delivery & Language Use	Unique words normalized by total word duration
27. Tpsecutt	Fluency & vocab. diversity	Delivery & Language Use	Unique words normalized by speech duration
28. Amscore	Pronunciation	Delivery	Acoustic model score; compares the pronunciation of nonnative speech to a reference pronunciation model (which models the probabilities of sequence of phonemes)
29. Lmscore	Grammatical accuracy	Language Use	Language model score; compares the language of nonnative speech to a reference language model (which models the probabilities of sequence of words)

Note. Mean deviation is computed as the mean of the absolute differences between feature values and the mean of feature values. The terms *pauses* and *silences* refer to the same thing. In all cases where the denominator would be zero, the respective value of a feature or component of a feature is also set to zero.

The automated features in Table 6 represent partially four aspects of the TOEFL iBT Speaking rubric: (a) fluency, (b) pronunciation, (c) vocabulary diversity, and (d) grammatical accuracy. Both fluency and pronunciation are related to the Delivery dimension of the rubric; vocabulary diversity and grammatical accuracy indicate the Language Use dimension.

Fluency Features and the Speaking Construct

There is a rich body of literature in second language learning examining how fluency varies for speakers of different proficiency levels (Deschamps, 1980; Raupach, 1980), how fluency is improved with language learning intervention or development (Dechert, 1980; Hansen, Gardner, & Pollard, 1998; Towell, 1987), or how some temporal variables are related to human judgments of fluency (Cucchiarini, Strik, & Boves, 2002; Freed, 1995; Lennon, 1990; Riggenbach, 1991). Some key, observable indicators of fluency have been identified, which include rate of speech, pauses, and length of runs between pauses.

As reviewed in Wood (2001), speech rate (wpsec) or articulation rate (phoneme per second) has been established as a major indicator of fluency. Greater speech or articulation rates are usually associated with more advanced speakers, although Munro and Derwing (2001) have found a curvilinear relationship between speech rate and overall comprehensibility. Their finding suggests an optimal speech rate, above which nonnative speech may be perceived as less comprehensible.

Two major aspects of pauses have attracted attention from the field: (a) duration and frequency of pauses and (b) location of pauses. A lower ratio of pause time to speech time and a lower relative frequency of unfilled pauses have generally been found to characterize speakers who are rated as more fluent (Lennon, 1984; Riggenbach, 1991). The location of pauses indicates the pause structure of speech. Research has shown that pausing at sentence or clause junctures or between meaning groups within a clause is related to perceived fluency (Freed, 1995; Riggenbach, 1991). When a speaker’s lexical, grammatical, or phonological encoding is less automatic or efficient, pausing in the middle of integral syntactic or meaning components occurs, which compromises the overall fluency and obscures meaning.

In addition to speech rate and pause time and structure, the mean length of runs between pauses has also found sufficient empirical support as an important indicator of fluency (Moehle, 1984). The more automatized, formulaic chunks of language a speaker has in store, the longer that speaker’s mean runs between pauses may be.

Conversation fillers (i.e., disfluencies) such as *um, er, uh, ah, okay, you see, I mean, you know, well*, and *so* are used by native speakers to fill silences, which in turn may make their speech sound more natural and fluent. Two factors may impact the effectiveness of communication when nonnative speakers use fillers: (a) To what extent are the fillers native like, and (b) how frequently are they used? Fillers that are not native like may be distracting to listeners. Excessive use of fillers may make one's speech sound less smooth. Although the use of fillers could be a useful feature, because our current speech recognizer does not identify fillers as well as human transcribers do, any filler-related automated features are not reliably identified at the moment.

Among the automated features that are related to fluency, wpsec (Feature 8) indicates the speech rate. We did not compute a speech articulation rate variable, although speech rate and articulation rate should be highly correlated. Features 12–22 are various measures for pauses. Among the pause measures, standard deviations or mean deviations of pause or long pause durations demonstrate whether the pauses are of varying or similar lengths. Features 6 and 7 indicate the mean length of runs in seconds and words between .20 second or longer pauses. We have not computed features that indicate the location of pauses yet. We would need a robust phrase chunker that can chunk utterances into phrases reliably. This is planned for future work.

Pronunciation Features and the Speaking Construct

The segmental elements of speech refer to the individual phonemes of the language. Pronunciation concerns how individual phonemes are produced. Along with goodness of prosody and grammatical errors, phonemic errors is one of the key factors that impact the intelligibility, perceived comprehensibility, and accentedness of speech (Derwing & Munro, 1997).

The quality of pronunciation is often judged in terms of how much listener effort is required to understand a speaker and to what extent phonemic errors interfere with meaning. The more difficulty a listener perceives in trying to understand a speaker (the more listener effort required), the more incomprehensible the speaker may be perceived. The more a speaker's errors obscure meaning, the greater difficulty a rater may have in understanding the speaker. We have extracted a pronunciation feature (amscore, Feature 28) to indicate the pronunciation quality of a speaker as compared to a reference pronunciation model.

Lexical Diversity Features and the Speaking Construct

Along with sophistication and precision of vocabulary, lexical diversity, or range of vocabulary, has often been used in the rubrics for speaking tests to define the quality of a speaker's vocabulary use (Read & Nation, 2004). The more repetitive a speaker's vocabulary is, the less likely that the speaker can express his or her ideas precisely.

The traditional type token ratio (number of unique words divided by number of words) has been criticized because of its sensitivity to length. Malvern and Richards (2002) proposed the D measure that fits a curve of type token ratios based on many different random samples of words in a text. They have claimed that this measure overcomes the disadvantages of the type token ratio because it is independent of sample size and it takes into account both long-distance and short-distance repetition by taking many random samples from a text. However, since the TOEFL speaking responses are extremely short, it may not be feasible to compute this D measure, which requires taking multiple samples of words. We have used an alternative way to control for length: types of words divided by the total length of speech (tpsecutt) or length of speech without pauses and disfluencies (tpsec). These two measures depend on both how fast a speaker talks and how many different types of words he or she uses in a given unit time. Therefore, we have categorized tssec and tspsectutt as both vocabulary diversity and fluency features. Tpsecutt measures fluency to a greater degree than tpsec.

Grammatical Accuracy Features and the Speaking Construct

Grammatical competence is defined as one of the four components of communicative competence in Canale and Swain (1980), in addition to discourse competence, sociolinguistic competence, and strategic competence. Grammatical accuracy in speaking rubrics is usually operationalized as the extent to which grammatical errors interfere with meaning. We currently have a grammatical accuracy feature that indicates the extent to which word sequences in a response conform to a reference grammar model that models the probability of different word strings (lmscore, Feature 29).

Design of the Study to Determine Which Features Will Be Adopted

Strategy and Expectations for Feature Use

In evaluating and selecting the final features for developing the scoring model, we considered both the construct representation of each feature (its linkage to the rubric and its conceptual overlap with other features) and its empirical performance, as indicated by the strength of its relationship with the human scores. Two basic principles were adopted in the feature selection process. First, we aimed to target as broad of a construct as possible with the final set of features we determined. Second, the substantive meaning of a feature was given more weight than its empirical correlation with the human scores in feature evaluation and selection.

Content representation. A CAC was convened that consisted of five assessment specialists with extensive experience in developing or rating speaking assessments. Two of them are intimately familiar with the TOEFL speaking scoring rubrics and are responsible for training and monitoring the TOEFL iBT Speaking test raters. This committee was charged with the task of reviewing the candidate speech features and the scoring models to make sure that the features are reasonable representations of the construct of speech and that the scoring models are substantively meaningful.

The automated scoring project team met with the CAC regularly. The following steps were followed to review and evaluate the features:

- The project team discussed the conceptual meaning and the computation of each feature with the CAC.
- The CAC reviewed the substantive meaning of all candidate features and suggested modifications of existing features or additional features to be computed.
- The project team discussed the basic statistics on all features with the CAC (descriptions of features, correlations of features with human scores, correlations among features, and differences in feature values across score levels).
- The CAC suggested modifications of features or additional features to be computed based on the design and empirical performance of the features.
- The CAC reviewed the modified or new features; this process was repeated until the committee members were comfortable with the changes.
- Using a formal rating form (see Appendix A), committee members independently rated how well each feature was linked to the rubric and how well the combined set of features represented the rubric.
- The CAC committee discussed their ratings as a whole group and modified their ratings, if necessary.

While selecting the features to be used in building the scoring models, the overlap among the features was also considered, in addition to the linkage of each feature to the rubric. The intercorrelations of a feature with other features as well as its conceptual overlap with them were taken into account while evaluating the unique contribution of a feature to the construct.

Empirical performance. As discussed above, the correlations of the features with the human holistic scores were also made available to the CAC members to help them evaluate the features. However, this information was provided only after they had a chance to review thoroughly the substantive meaning of the features. The statistical properties of the features provided the content specialists with another perspective on the features. Although the judgmental evaluations of the features supported a proper representation of the construct with all the features combined, examining the correlations of the features with the human scores ensured that the features were useful in predicting the human scores.

Data Used in the Study

We made use of the TPO data in training the speech recognizer and in evaluating the speech features. The rec-train set was used for recognizer training, whereas the sm-train set was used to evaluate the statistical properties of features before model building and to parameterize the scoring models. The sm-eval set was used to calculate the final evaluation statistics summarizing each model's agreement with human raters.

Analytic Design and Analyses

The speech recognizer: Acoustic properties of the data. The audio signals are sampled with a sampling rate of 22,050 Hz, mono, 16-bit resolution. The signal then is compressed into the Windows Media Format (.wma files; compression is not loss free) and has to be downsampled to 11,025 Hz and converted to standard PCM (.wav files) by means of the freely available program FFmpeg (n.d.), as our recognizer cannot handle .wma files. Although the wma compression not being loss free may have a slight adverse effect on speech recognition, the effect of the downsampling should be less noticeable, since the major speech events happen below 5 kHz, which is well covered with a sampling rate of 11 kHz (covers 0–5,500 Hz range).

The speech recognizer: Training and adaptation of the speech recognizer. For the experiments in this report as well as for the operational TPO SpeechRater engine, we used a recognizer by Multimodal Technologies, Inc., specifically trained on the TPO data. Multimodal bootstrapped this recognizer by using an existing 11 kHz recognizer trained on a large set of transcribed speech of native speakers of American English. The AM was then adapted to the TPO data (rec-train, about 2,000 speech samples of 45–60 seconds each) and the recognizer was retrained. For the LM, Multimodal used responses to both prototype and official TOEFL speaking tasks that were captured in various ways (i.e., Internet and the Interactive Voice Response system) as well as data from the Linguistic Data Consortium (Fiscus, Garofolo, Praybocki, Fisher, & Pallett, 1997). A trigram model with absolute discounting was used (Manning & Schuetze, 1999). The vocabulary size is 11,019 unique words, and the pronunciation dictionary contains 11,823 entries, 804 of which are pronunciation variants. The out-of-vocabulary rate on a token basis on the transcribed part of the sm-train data set (645 files, about 67,000 words not including fillers) was measured to be 0.8% This out-of-vocabulary number means that 0.8% of the words (tokens) in the transcribed sm-train data set were not found in the recognizer's pronunciation dictionary. The phoneme set contains 40 different regular phonemes, as well as some special-purpose phonemes for silence and filled pauses ("uh," "um").

The speech recognizer: Speech decoding process. Using the speech recognizer described above, we decoded both the transcribed portions of the sm-train and sm-eval sets in three different speed-versus-accuracy (SvA) settings: (a) SvA = 0.0 (*fast, low accuracy*), (b) SvA = 0.5 (*balanced*), and (c) SvA = 1.0 (*high accuracy, slow*). The rationale behind this was to explore the trade-off between speed and accuracy of the recognizer, as we desired to have a short response time for TPO in its operational setting without losing too much word accuracy.

Table 7 provides the results both in terms of word accuracy and processing time. As shown in Table 7, SvA = 1.0 yields only a small gain in word accuracy at the cost of a more than six-fold increase in decoding time compared to SvA = 0.5. For that reason, we decided to set SvA to 0.5 in the operational system. When using smaller SvA values (e.g., SvA = 0.0), word accuracies are markedly lower; however, some preliminary evidence seems to indicate that despite the lower word accuracy, the effects on the accuracy of the scoring model are minimal. We will consider lowering the value of SvA in future versions of SpeechRater in the interest of a shorter processing time but are aware that additional features in the grammar, vocabulary, or content domains may be more sensitive to lower word accuracies.

Table 7

Word Accuracy and Decoding Speed on Scoring-Model Training and Evaluation Sets

Speed-vs.-accuracy data set	*n*	Word accuracy[a]	Average decoding time in seconds per speech sample	Decoding time *SD*
0.0 – train	645	41.0%	10.1	2.6
0.0 – test	395	43.1%	10.4	2.3
0.5 – train	645	48.9%	54.9	16.8
0.5 – test	395	51.4%	56.9	14.7
1.0 – train	645	50.5%	366.2	208.1
1.0 – test	395	53.3%	366.0	208.2

Note. Data set is number of transcribed samples (only applies to word accuracies).

[a] The word accuracy is computed as in previous research (Zechner et al., 2007): word accuracy = 0.5 * 100 * (C / (C + S + D) + C / (C + S + I)), with C = correct, S = substitutions, D = deletions, I = insertions, as a result of a string alignment with Levenshtein distance. See *Towards an Understanding of the Role of Speech Recognition in Non-Native Speech Assessment,* by K. Zechner, I. I. Bejar, and R. Hemat, 2007, Princeton, NJ: ETS; “Binary Codes Capable of Correcting Deletions, Insertions, and Reversals,” by V. I. Levenshtein, *Soviet Physics Doklady, 10*.

The speech recognizer: Speech recognizer output. The recognizer writes one line of output for every word recognized. It consists of an utterance label (usually composed of speaker, item, and utterance ID), the start time and word duration in seconds, the word itself, and a confidence score (between 0.0 and 1.0). This confidence score is based on posterior recognition probabilities but is not to be interpreted as the probability of a word being correctly recognized; it is just a weak indicator (see Appendix B for a sample recognizer output).

Furthermore, at the beginning a HEADER line contains the sample ID. At the end a TRAILER line contains the AM and LM scores as reported by the recognizer after recognition is completed.

In addition, every utterance has a pair of prosodic feature vectors, one for power (UTT-Power) and one for pitch (UTT-SmoothPitch). The values in these vectors are various moments of the features over the period of the whole utterance as well as minimum and maximum values in that segment. They are only used for the filtering model, which decides which responses to assign a TD or 0 score and which to route to the main scoring model.

Feature meaning and performance with human scores. Having provided input on the features following the process described earlier, the CAC made formal evaluations of the construct linkage and coverage of the features. Using the formal rating form in Appendix A, they first rated independently how well each feature was linked to the rubric and represented the feature class (e.g., fluency) and the dimension (Delivery, Language Use, and Topic Development) and how well the combined set of features represented the rubric. Then, they discussed their ratings as a whole group and adjusted their ratings, if necessary. Their evaluations of the features provided a basis for us to select the final features for the scoring models. A total of 13 features were selected based on their substantive meaning only.

Then, the intercorrelations among these features were checked. If two features were correlated at .90 or higher, one of them was excluded based upon linkage to the construct, conceptual overlap with other existing features, and strengths of relationships with the human scores. This process eliminated two features, wpsecutt and silpsec. Wpsecutt was removed because it was one of the many fluency features, whereas amscore was the only pronunciation feature available. Silpwd was selected rather than silpsec because both indicate the same aspect of fluency conceptually, and the former was rated as more meaningful by the CAC and had a slightly higher correlation with the human scores as well.

The correlations of these features with human scores along with the average ratings of these features on Questions 1–3 in Appendix A by the CAC members are included in Table 8. For all the features, the ratings on Question 1, which asked how well each feature is linked to a key dimension in the rubric, were 4 or higher on a 6-point scale. This attests to the meaningfulness of the features as perceived by the content specialists. The CAC members' ratings on Questions 2 and 3, which targeted the construct coverage of the features of a particular feature class or dimension, were also generally high. This offers additional evidence that these were important features that covered a key feature class or dimension well.

Table 8

Final Set of Features Used in Building the Scoring Models

Feature[a]	Feature class	Corr. with human holistic scores [b]	Average ratings by CAC by question (Q)		
			Q1	Q2	Q3
13. Silpwd	Fluency	-.294	4.6	4.4	4.4
15. Silmean	Fluency	-.282	4.7	4.4	4.4
8. Wpsec	Fluency	.449	5.1	4.5	4.3
26. Tpsec	Fluency & vocabulary	.296	5.1	4.5	5.1
6. Wdpchk	Fluency	.106	5.6	5.2	4.8
19. Longpwd	Fluency	-.327	4.2	3.6	3.6
18. Longpmn	Fluency	-.204	4.5	3.8	3.8
28. Amscore	Pronunciation	-.445	4.7	4.2	4.2
29. Lmscore	Grammar	-.295	4.4	4.2	3.8
27. Tpsecutt	Fluency & vocabulary	.408	5.4	5.1	5.1
11. Wdpchkmeandev	Fluency	.097	4.2	3.6	3.8
9. Wpsecutt[c d]	Fluency	.490	5.4	4.6	4.5
14. Silpsec[c e]	Fluency	-.149	3.1	3.1	3.1

Note. CAC = Content Advisory Committee.

[a] The feature numbers are consistent with those in Table 6. [b] These were correlations before some features were transformed. [c] Removed due to high correlations with other features. [d] Correlation with amscore: .94. [e] Correlation with silpwd: .93.

6. Development and Validation of the Scoring Method

This section describes the development and evaluation of different methods of spoken response scoring and the rationale we used to select the final scoring model for the SpeechRater system. The evaluation addresses the appropriateness of the scoring models to the construct as well as the empirical performance of the scoring models in relation to human scores (both in terms of agreement and in terms of reliability). Therefore, the results presented in this section constitute evidence for the Explanation, Evaluation, and Generalization inferences.

Standards for Evaluating the Scoring Models

Our goals are to develop a scoring model that meets our expectations for technical standards and is a reasonable representation of the speaking construct. The standards we used to evaluate the substantive meaning and the technical quality of the scoring models are described below.

Evaluating the Construct Representation

In the evaluation of the construct representation of the features, the following factors were considered: (a) the extent to which the features in the scoring models are linked to and cover the construct and (b) the extent to which the way the features are combined to produce scores captures the expected relationships between the features and the speaking scores. Furthermore, in considering models based on multiple regression and classification and regression trees (CART), we were willing to sacrifice a bit of accuracy in order to ensure appropriate construct representation.

Evaluating Technical Quality

Our evaluation of the technical quality of the scoring models considered below focuses on four aspects:

- Agreement of automated scores with human scores,
- Degradation of human–automated score agreement from human–human agreement,
- Mean score differences between automated and human scores, and
- Generalizability of automated scores across tasks.

The primary measures we used to assess the level of agreement between our models' predicted scores and the scores assigned by humans were the coefficient of correlation and the root mean squared error (RMSE). A secondary criterion that we report is the weighted κ (Cohen, 1968). (We used the quadratic weighting scheme, so that the penalty associated with a wrong score increases with the square of the difference between the human and predicted scores.) We used the weighted κ only as a secondary criterion for two reasons. First, the weighted κ is computed in terms of rounded scores and therefore is based on incomplete information about the scoring model's prediction, especially at the task score level. Second, the weighted κ increases with the standard deviation of the predicted scores, and we do not want to encourage scoring models with a high variance.

Finally, we generally report the exact accuracy and *exact + adjacent* accuracy of the scoring models, which indicate, respectively, how often the predicted (rounded) score is exactly the same as the human-assigned score and how often the predicted score is off by no more than one score point. These last two statistics are poor measures of overall model quality, but we provide them because they have been used widely in other work on automated scoring.

When possible (i.e., when the model to be considered is a regression model), we report the coefficient of correlation and the RMSE using unrounded predictions of scores. Generally, these better reflect the model's consistency with human ratings. Rounding the scores before calculating agreement metrics loses a great deal of information about the model's discriminative power. However, we also provide these metrics calculated using rounded scores, for all models, so that we have standard of comparison that applies to both regression-based and classification-based methods.

As useful summary statistics, we report the mean and standard deviation of the models' predicted scores and of the human scores assigned to each set of responses. These serve to measure any bias or shrinkage the models might exhibit. In addition, we report the generalizability of the models' predicted scores across different tasks, as this is an important piece of information to consider in evaluating an automated scoring system.

It is important to emphasize that whether a particular level of prediction accuracy is acceptable depends on the intended use of the scores. Although degradation from human–human agreement is important to look at, the absolute level of performance determines the ultimate practical value of this system in supporting the intended use of the whole practice test.

In the development of these evaluation criteria and their application to the scores produced by SpeechRater, we were assisted by a panel of psychometricians and measurement statisticians who were updated frequently on the status of the research. These experts comprised our Technical Advisory Committee (TAC) and played a comparable role in their oversight of the measurement issues encountered in the course of the project to that played by the CAC regarding construct issues. The TAC members had extensive previous experience with automated constructed-response scoring technologies, having supervised a number of applications of e-rater (Attali & Burstein, 2006) for the scoring of writing. In addition, they had prior experience with the spoken-response item types used on the TOEFL iBT. Therefore, their advice was well informed about measurement issues arising independently in the domains of automated scoring and spoken-response scoring, and they were well prepared to consider this area of intersection between the two.

Design of the Study to Determine What Scoring Model Would Be Applied

Models Under Consideration: Characteristics and Expected Strengths and Weaknesses

In this study, two methods for building scoring models were evaluated: multiple regression and CART (Breiman, Jerome, Olshen, & Stone, 1984). Other model types, such as Bayes networks (Pearl, 1988) and logistic regression, are potentially worthy of investigation for such an application, but the scope of this study was restricted to more well-known models for the initial release of SpeechRater. Multiple regression has the advantage of making predictions on a real-valued scale and does not discard information about centrality of class membership by forcing predicted scores to be integers. Also, being commonly known and applied in the social sciences, multiple regression is a method that is more readily understandable to potential users of scores from SpeechRater than less familiar statistical methods. As a parametric method, it is also more stable and statistically flexible. The CART method, on the other hand, hold the promise of a model structure that is more congruent with the way in which trained raters make their judgments at the task level, since it does not impose the same model structure over the entire score scale.

Linear regression. The multiple regression scoring model calculates a predicted score for a response as a linear combination of feature values. To be concrete, the score is calculated according to the equation

$$Score = \sum_i \alpha_i f_i + \beta \, .$$

In this equation, i is the index of each feature in the model, the f_i are the feature values, and the α_i are the weights associated with each feature. β is a constant intercept term.

Modeling the score as a weighted sum of feature values means that the features must be defined in such a way that the relationship between features and scores is the same throughout the entire range of the score scale. This also means that the predicted score may be a real (noninteger) number, which must be rounded to be used as a score that can be reported to the examinee. The fact that this score is a real number, however, is useful in doing score aggregation, because more information about an examinee’s performance on each task is retained.

CART. Classification trees are hierarchical, sequential classification structures that recursively partition the observations into different classes. At each decision node, the variable that best can classify the cases into distinct classes at a certain value is selected to perform the partition. For example, as illustrated in Figure 5, if the rate of speech is less than 3 wpsec, a response goes to the left node (representing the lower score class) and to the right node (representing the higher score class) if otherwise. Then, each of the child nodes may be further partitioned into two more nodes down the tree. The process is repeated until a *terminal node* is reached. In a nutshell, classification trees yield a set of if–then logical (split) conditions that permit the classification of cases.

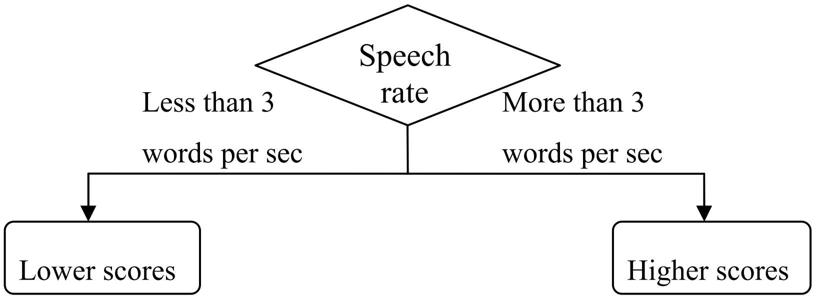


***Figure 5.* An illustrative example of a classification tree.**

Although a few computer programs implement this methodology, the CART software (Steinberg & Colla, 1997) implements the original methodology developed by Breiman et al. (1984). CART analyses are typically conducted in three steps:

1. In tree growing, the maximum tree is grown on the training sample.
2. In tree pruning, the maximum tree is pruned to produce a sequence of nested trees, and error rates are obtained for each tree using a cross-validation sample or a testing sample.
3. In optimal tree selection, the best tree is identified that represents a balance between the error rate and tree complexity.

In CART, a few methods are essential for refining the analyses for special circumstances: the use of prior probabilities, the use of different weights (costs) on different misclassification errors, and the use of different splitting rules. When the probabilities of different classes are known in the population but the sample used to train the data is not representative of the population, it may be important to use the priors of the population, since this will affect the estimation of error rates. Priors can also be used to avoid misclassifying certain classes that are deemed important. By increasing the prior probabilities of these classes in the population, trees can be grown that misclassify them less often. Either a higher probability or a higher cost can be put on an important class to improve its classification. However, misclassification costs can be used to define the costs of specific errors (e.g. the cost of classifying a score of 1 as 3 or a score of 4 as 2). This level of control is not available by manipulating priors. In addition to manipulating priors and costs, CART also offers a variety of splitting rules, including Gini, twoing, ordered twoing, class probability, forcing splits by user, and linear combination splits (see Steinberg & Colla, 1997, for a detailed description of these splitting rules). It is usually preferable to explore different splitting rules to find out which one provides the best classification results in a particular problem.

An advantage of classification trees is that they do not assume that the underlying relationships between the predictor variables and the predicted classes are linear; do not follow some specific, nonlinear link function; and are not monotonic in nature. For example, holistic speaking scores could be positively related to speech rate if the rate is less than 5 wpsec but negatively related if the rate is greater than that. That is to say, the tree could have multiple splits based on the same variable, revealing a nonmonotonic relationship between the variable and the predicted classes. Moreover, a variable that does not discriminate well in the higher score classes can be used in classifying lower score classes without impacting the prediction of the higher score classes. These characteristics of classification trees contrast with other classification or prediction techniques, which use *all* of the important variables for classifying or predicting each case. Since the distinguishing speech features for different score classes may be different and the relationship between a speech feature and the speaking score may not be linear, conceptually, classification trees appear to be a suitable technique for classifying score classes based on scores from the TOEFL iBT Speaking Practice test. In addition, different patterns of strengths and weaknesses in different aspects of speech may lead to the same score class. This is also compatible with a feature of classification trees that different sets of decision rules may result in the same score class.

Although complex mathematical computations are used in growing the trees, the actual application of the tree is performed using a sequence of simple and easily understood decision rules that are transparent to content specialists. The classification mechanism is intuitively appealing. Thus, this technique is amenable to incorporating content specialists' input in evaluating and refining the tree structure that best represents expert raters' decision processes.

Data Used in the Study

In building and evaluating the scoring models described below, we made use of both the TPO data and the field study data (described in detail in Section 6). For the field study data set, the sm-train set was used for model building and the **s**m-eval set for evaluation. For the TPO data set, the sm-train set was used for model building. For evaluation we used both the sm-eval set by itself as well as the sm-eval set combined with the rec-train set (see below).

Analytic Design and Analyses

Statistical transformations and outlier processing. One complication of the multiple regression approach is the features we have developed for speech scoring may not conform to model assumptions, notably, the assumption of a linear relationship between the features and the score and the assumption that the error term in the regression equation is normally distributed. To address this possibility, we examined the distribution of each of our features and considered transformations of the features that might improve the correlation between the feature and the item score to be predicted, as well as making the feature's distribution more normal.

In determining whether a given feature ought to be transformed before being used in the regression model, we used two sources of information: normality information of the feature itself and correlation between the feature and the human task-level score. Only transformations that resulted in substantial improvements in normality or increases in correlations were used.

We used quantile-quantile (Q-Q) plots to determine how far the feature's distribution diverged from normality. Figure 6 shows such Q-Q plots for the amscore feature, before and after the inverse transformation. The plot on the left shows that the untransformed feature diverges from normality with a positive skew, whereas the plot on the right shows that the inverse transformation yields a feature whose distribution almost perfectly matches a Gaussian, with the Q-Q plot being almost a straight line.

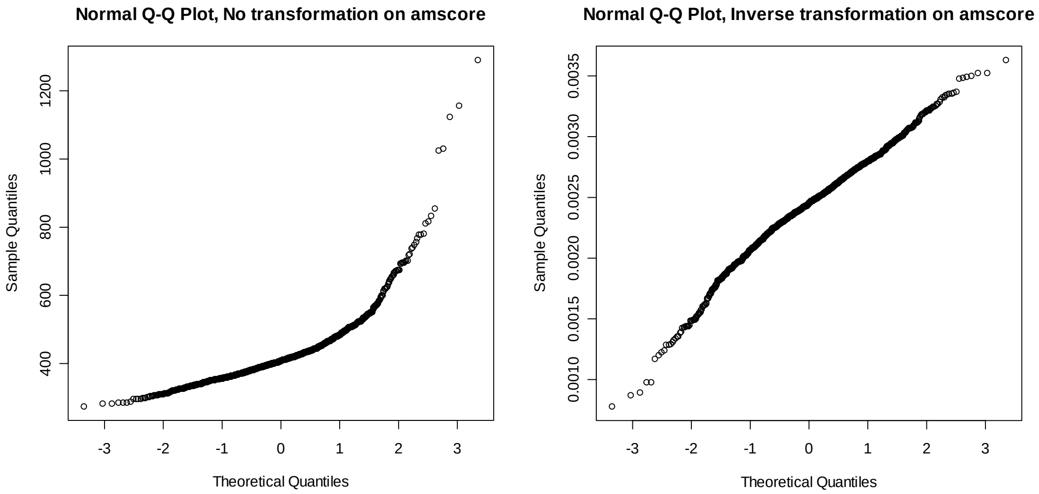


Figure 6. **Quantile-quantile (Q-Q) plots for the feature amscore.**

Table 9 shows the transformation performed on each feature and the changes in correlations before and after the transformation. Correlations for three of the four features improved considerably after the transformations, with one staying almost the same. The distributions of all of the four features were also much more normal after the transformations.

Table 9

Changes in Correlation Before and After the Transformation

		Correlations with human scores	
Feature	Transformation performed	Original	Transformed
Wdpchk	Natural log (wdpchk + 1)	.106	.222
Amscore	Inverse	-.445	.510
Lmscore	Inverse	-.295	.282
Wdpchkmeandev	Inverse	.097	-.248

Since multiple regression is also susceptible to outliers, we also examined the feature distributions for outliers, which show up as deviations in the tails of our Q-Q plots. Ultimately, we decided that a cutoff value of 4 standard deviations from the mean best isolated the outliers in our training data, and we mapped all feature values outside this range to the maximum or minimum values allowed. Because classification trees are robust to outliers and do not assume normality of the data (Steinberg & Colla, 1995), no variable transformations or outlier processing were performed on the features used by the CART models.

Model building: Multiple regression. Our aim in developing the SpeechRater multiple regression model was to produce a model with high agreement with human raters, but also to structure the model so that its use of our predictive features is in conformance with our understanding of the speaking construct. Toward this end, we restructured the regression equation shown above (and repeated here for the sake of clarity) to use fixed feature weights instead of empirically determined weights:

$$Score = \sum_i \alpha_i f_i + \beta \,.$$

This original equation had a free parameter *α* associated with each scoring feature. The new equation still has a parameter *α'* associated with each feature, but these parameters are not allowed to vary in the optimization of the model for a given training set. The only two parameters that need to be learned from the data are the slope parameter *μ* and the intercept *β:*

$$Score = \mu \sum_i \alpha_i' f_i + \beta \,.$$

In essence, the training for a model of this form reduces to a rescaling of the linear regression function determined by the fixed weights α_i'. Note that a scoring function in this form always can be mapped back to a format in which each feature has a weight parameter, and the slope parameter *μ* is not factored out, using the mapping function $\alpha_i = \mu\alpha_i'$.

The feature-specific weights α_i' were specified in consultation with the CAC, which, as discussed above, was a group of content-area specialists convened to ensure the construct appropriateness of our scoring model design. Note also that we standardized the feature values (to zero mean and unit variance) so that the CAC weights assigned were comparable across all features. The standardization parameters (the mean and variance of the feature as observed in the training data) were retained for scaling of the features in the test samples as well.

The CAC agreed on the use of a model with the features amscore, wpsec, tpsecutt, wdpchk, and lmscore. This set of features was deemed to provide the widest range of coverage of the different aspects of the speaking construct and could be weighted in such a way that the relative importance of each of these measures was represented. The weights for each feature were set by the research team with prior knowledge of the statistical properties of the features, their correlations with human scores, and the weights assigned to each feature by least-squares optimization. This weighting scheme was discussed and ultimately endorsed by the CAC. Table 10 provides the feature class and dimension represented by each of the features used, together with their weights in the regression model.

Table 10

Features Used in Content Advisory Committee(CAC) Regression Model

Feature	Weight	Feature class	Dimension
Amscore	4	Pronunciation	Delivery
Wpsec	2	Fluency	Delivery
Tpsecutt	2	Vocabulary, fluency	Delivery & Language Use
Wdpchk	1	Fluency	Delivery
Lmscore	1	Grammar	Language Use

Other models with different features and weightings were considered, but their agreement with human raters was not significantly different from the CAC regression model discussed here, and they had inferior construct representation. For these reasons, we focus only on regression models that use the CAC feature set shown in Table 10.

We fixed the weights (α') of the standardized features to the CAC-defined values shown in Table 10 and trained the model using the TPO sm-train data as described earlier. This involved setting the model slope parameter μ and intercept β to minimize the least squares error on this training data.

In addition, we built a model, the *equal weights model*, that uses the same feature set (shown in Table 10) but assigns them equal weights, rather than the expert weights we had used previously. The third model we developed, the *optimal weights model*, assigns least squares optimal weights to each feature, rather than setting them at fixed values.

One shortcoming of just using the TPO sm-eval set for evaluation is that it does not allow a direct estimation of the correlation of predicted scores with human-assigned scores on a full complement of six tasks, which is the level of the score we wish to report (since there are six tasks in the TOEFL iBT Practice Speaking test). There were only 58 candidates with complete sets of six task scores in this evaluation set. To address this deficiency, we performed an additional evaluation run on the combined data from the TPO sm-eval and rec-train sets. This combined set of evaluation data contained many more (308) complete sets of six tasks per candidate than the sm-eval set alone.

Strictly speaking, however, there is a methodological issue with doing the evaluation this way. Since the data from the rec-eval set were used to train the speech recognizer, some of the learning from this stage (relative probabilities of word sequences and pronunciation variants) might cause the scoring model to perform uncharacteristically on this particular set of data. In practice, however, this seems unlikely, given that our feature set abstracts away from the actual hypothesized word sequence returned by the recognizer. Although the lmscore and amscore features do use information about the internal state of the recognizer, and therefore could be affected by the use of a particular response in recognizer training, we expect this effect to be small.

Model building: CART trees. CART 5.0 (Steinberg & Colla, 1997) was used to build the classification trees. We explored different model configurations using different combinations of priors and splitting rules. For each combination, a 10-fold cross-validation was conducted. In each set of 10-fold cross-validation, a tree was first grown on the entire sm-train sample, yielding a sequence of trees for which error rates could be computed. Then, the sm-train set was divided into 10 subsets of equal sizes, stratified on the dependent variable. In each cross-validation run, one subset was used as the testing sample, with the remaining 9 subsets as the training sample. This process was repeated 10 times. After the completion of the 10 runs, the error counts from each of the 10 test samples were summed to obtain the overall error count for each subtree in the whole-sample tree sequence (Steinberg & Colla, 1997). Subsequently, the optimal subtree was identified: a relatively small tree with the highest or near-highest agreement with the human scores (weighted kappa) on the cross-validation sample.

Substantive meaning of models. To evaluate the construct representation of each scoring model, four CAC committee members provided overall ratings of the construct representation of each model using a formal evaluation form (Appendix C). They considered the relevance and coverage of the features present in the model as well as the meaningfulness of the contribution of features to scores. The first two questions asked how well the features present in the model represented the TOEFL iBT speaking rubric and how well the model captured the relationships between the automated features and the speaking construct. The third question asked their opinions on the extent to which the model was consistent with the decision-making processes that human raters use to derive a holistic score.

Model performance in terms of agreement with human scores. As discussed in the section on the standards used to evaluate the technical quality of the scoring models, we produced various statistics that indicate the correspondence of the scores predicted by the scoring models with the human scores.

Model performance in terms of score generalizability across tasks. Generalizability studies were conducted on the scores for the sm-eval and rec-train sets estimated using multiple regression and CART, respectively. The phi coefficients for scores across six tasks were computed for the two scoring models and compared to that of the single human scores summed across six tasks. The phi coefficient indicates the dependability of scores when absolute decisions are of concern, that is, when the absolute levels of scores, rather than the rank ordering of them, are of interest.

Results

Multiple regression. The result of applying this scoring model to the sm-eval set is shown in the center column of Table 11. In Table 11, the agreement results are broken down into three sections. In the top third of the table, we show the agreement between SpeechRater's predicted scores and the scores assigned by human raters to a single task. In the second third, we provide the same information for aggregated pairs of scores (sums of scores on two tasks by the same examinee). The bottom third of the table provides the same information for sets of three scores. Because of the way the data were partitioned for this project, there were not enough examinees with six complete tasks (one full test form) in the sm-eval set for us to perform this evaluation on a full set of six aggregated tasks.

The most important statistic for evaluating the quality of this scoring model is the correlation of the predicted scores with the human-assigned scores, which ranges from .49 for single items to .56 for sets of three tasks. (We report correlations for unrounded scores, as these allow for better extrapolation to the correlations to be expected on sets of six tasks, the scores we

Table 11

Performance of Different Weighting Schemes Using CAC Feature Set on TPO Scoring-Model Evaluation Set Data

<table>
<tr><th rowspan="2">Model</th><th colspan="15">Multiple regression</th></tr>
<tr><th colspan="5">Equal weights</th><th colspan="5">CAC weights</th><th colspan="5">Optimal weights</th></tr>
<tr><td colspan="16">Single scores (<i>N</i> = 520)</td></tr>
<tr><td>Quadratic-weighted κ</td><td colspan="5">0.30</td><td colspan="5">0.32</td><td colspan="5">0.35</td></tr>
<tr><td>Exact agreement</td><td colspan="5">59.0%</td><td colspan="5">59.2%</td><td colspan="5">61.3%</td></tr>
<tr><td>Exact + adjacent agreement</td><td colspan="5">98.5%</td><td colspan="5">98.8%</td><td colspan="5">98.7%</td></tr>
<tr><td>Mean (<i>SD</i>) of predicted score</td><td colspan="5">2.78 (.32)</td><td colspan="5">2.78 (.33)</td><td colspan="5">2.73 (.31)</td></tr>
<tr><td>RMSE (unrounded)</td><td colspan="5">0.61</td><td colspan="5">0.60</td><td colspan="5">0.61</td></tr>
<tr><td>Correlation (unrounded)</td><td colspan="5">0.46</td><td colspan="5">0.49</td><td colspan="5">0.47</td></tr>
<tr><td>RMSE (rounded)</td><td colspan="5">0.68</td><td colspan="5">0.67</td><td colspan="5">0.65</td></tr>
<tr><td>Correlation (rounded)</td><td colspan="5">0.35</td><td colspan="5">0.37</td><td colspan="5">0.40</td></tr>
<tr><td rowspan="5">Confusion matrix: <i>predicted</i> vs. human scores</td><td></td><td><i>1</i></td><td><i>2</i></td><td><i>3</i></td><td><i>4</i></td><td></td><td><i>1</i></td><td><i>2</i></td><td><i>3</i></td><td><i>4</i></td><td></td><td><i>1</i></td><td><i>2</i></td><td><i>3</i></td><td><i>4</i></td></tr>
<tr><td>1</td><td>0</td><td>13</td><td>5</td><td>0</td><td>1</td><td>0</td><td>13</td><td>5</td><td>0</td><td>1</td><td>1</td><td>12</td><td>5</td><td>0</td></tr>
<tr><td>2</td><td>0</td><td>52</td><td>106</td><td>1</td><td>2</td><td>0</td><td>54</td><td>105</td><td>0</td><td>2</td><td>0</td><td>63</td><td>96</td><td>0</td></tr>
<tr><td>3</td><td>0</td><td>31</td><td>254</td><td>4</td><td>3</td><td>0</td><td>34</td><td>252</td><td>3</td><td>3</td><td>0</td><td>35</td><td>253</td><td>1</td></tr>
<tr><td>4</td><td>0</td><td>2</td><td>51</td><td>1</td><td>4</td><td>0</td><td>1</td><td>51</td><td>2</td><td>4</td><td>0</td><td>2</td><td>50</td><td>2</td></tr>
<tr><td colspan="16">Paired scores (<i>N</i> = 232)</td></tr>
<tr><td>Weighted κ</td><td colspan="5">0.35</td><td colspan="5">0.40</td><td colspan="5">0.41</td></tr>
<tr><td>Mean (<i>SD</i>) of predicted scores</td><td colspan="5">5.56 (.57)</td><td colspan="5">5.55 (.58)</td><td colspan="5">5.47 (.53)</td></tr>
<tr><td>RMSE (unrounded)</td><td colspan="5">0.99</td><td colspan="5">0.97</td><td colspan="5">0.98</td></tr>
<tr><td>Correlation (unrounded)</td><td colspan="5">0.49</td><td colspan="5">0.52</td><td colspan="5">0.50</td></tr>
<tr><td>RMSE (rounded)</td><td colspan="5">1.05</td><td colspan="5">1.01</td><td colspan="5">1.00</td></tr>
<tr><td>Correlation (rounded)</td><td colspan="5">0.42</td><td colspan="5">0.47</td><td colspan="5">0.48</td></tr>
<tr><td colspan="16">Triples of scores (<i>N</i> = 163)</td></tr>
<tr><td>Weighted κ</td><td colspan="5">0.43</td><td colspan="5">0.44</td><td colspan="5">0.41</td></tr>
<tr><td>Mean (<i>SD</i>) of predicted scores</td><td colspan="5">8.39 (.79)</td><td colspan="5">8.38 (.83)</td><td colspan="5">8.25 (.77)</td></tr>
<tr><td>RMSE (unrounded)</td><td colspan="5">1.33</td><td colspan="5">1.30</td><td colspan="5">1.33</td></tr>
<tr><td>Correlation (unrounded)</td><td colspan="5">0.53</td><td colspan="5">0.56</td><td colspan="5">0.52</td></tr>
<tr><td>RMSE (rounded)</td><td colspan="5">1.35</td><td colspan="5">1.35</td><td colspan="5">1.37</td></tr>
<tr><td>Correlation (rounded)</td><td colspan="5">0.51</td><td colspan="5">0.51</td><td colspan="5">0.48</td></tr>
</table>

Note. CAC = Content Advisory Committee.

ultimately hope to report.) These correlations are not as strong as we might hope for; however, they must be evaluated taking account of the more limited variability in the scores in this sample than in the field study sample. Remember from the discussion of human agreement above that the correlation between the scores assigned by two human raters was only about .70 on sets of three items from this data set. That effectively sets an upper bound on the effectiveness of our model (and also makes it more attractive to find a replacement for human scoring, which is currently somewhat unstable).

Note also that there is less variation in SpeechRater's score estimates. The standard deviation of predicted scores is considerably lower than the standard deviation of human-assigned scores (0.33[1] vs. 0.69[2] for single items). This is partially due to the uneven distribution of task scores in the training data (with almost half the tasks receiving a score of 3) but also may have to do with inconsistency in the human scoring of these responses and with the limited range of construct coverage in our feature set.

Table 11 also provides results for two models related to the multiple regression scoring model already discussed, which had feature weights set by the CAC. The model shown in the first column is the equal weights model and the final column of the table has the results for the optimal weights model. (This optimal weights model differs primarily in assigning even more importance to the pronunciation feature amscore). The summary statistics for these three models are very similar, illustrating that the multiple regression model is not sensitive to small variations in the weights chosen (cf. Wainer, 1976). In fact, the regression model with the weights set by the CAC has the best correlation with human scores for sets of three tasks.

In order to perform an analysis to aggregate complete sets of six responses, we also tested our CAC regression model on a data set composed of the sm-eval set combined with the rec-train set. Although the rec-train set is not technically a pure unseen evaluation set, because it was used in the training of the speech recognizer, it was not used in the parameterization of the scoring model itself. Any scoring differences observed in this data set could be only a result of a difference in the recognizer's word error rate, which is unlikely to be a strong predictor of scoring accuracy, because our features are not dependent on a highly faithful recovery of the words in the response.

The results of the CAC regression model are shown in the second column of Table 12 for this combined data set of sm-eval and rec-train. These are generally in line with the results shown

Table 12

CAC Regression Model Performance on TPO Evaluation + Recognizer Training Set and on Field Study Data Set

Scoring-model evaluation set	TPO evaluation + recognizer training set					Field study test set				
Single scores	*N* = 2427					*N* = 1752				
Quadratic-weighted κ	0.325					0.510				
Exact agreement	57.8%					44.2%				
Exact + adjacent agreement	98.4%					95.1%				
Mean (*SD*) of predicted score	2.79 (.37)					2.45 (.61)				
RMSE (unrounded)	0.63					0.79				
Correlation (unrounded)	0.47					0.61				
RMSE (rounded)	0.69					0.84				
Correlation (rounded)	0.37					0.55				
Confusion matrix: *predicted* vs. human scores		*1*	*2*	*3*	*4*		*1*	*2*	*3*	*4*
	1	0	49	21	0	1	91	217	31	0
	2	0	254	451	4	2	25	310	215	3
	3	0	154	1,120	26	3	14	172	348	8
	4	0	15	305	28	4	1	36	255	26
Pairs of scores	*N* = 1,137					*N* = 854				
Weighted κ	0.45					0.58				
Mean (*SD*) of predicted scores	5.58 (.68)					4.92 (1.13)				
RMSE (unrounded)	1.04					1.36				
Correlation (unrounded)	0.53					0.66				
RMSE (rounded)	1.06					1.40				
Correlation (rounded)	0.50					0.64				
Triples of scores	*N* = 757					*N* = 555				
Weighted κ	0.48					0.61				
Mean (*SD*) of predicted scores	8.40 (.99)					7.43 (1.63)				
RMSE (unrounded)	1.40					1.90				
Correlation (unrounded)	0.56					0.68				
RMSE (rounded)	1.42					1.92				
Correlation (rounded)	0.54					0.67				
Sets of six scores	*N* = 308					*N* = 254				
Weighted κ	0.51					0.61				
Mean (*SD*) of predicted scores	16.84 (1.87)					15.13 (3.04)				
RMSE (unrounded)	2.48					3.57				
Correlation (unrounded)	0.57					0.68				
RMSE (rounded)	2.50					3.56				
Correlation (rounded)	0.57					0.68				

Note. CAC = Content Advisory Committee, TPO = TOEFL Practice Online.

above for the sm-eval set only, but with the addition of results for a complete set of six tasks. For this total raw score summed across six tasks, the correlation between predicted scores and human-assigned scores is .57. This is somewhat higher than we saw previously, for smaller sets of aggregated items, but still lower than we would like.

As a final evaluation, in order to determine how much the performance of our scoring model might be obscured by the variability in the human scores assigned to the TPO data, we applied the same regression model to the field study data set. Because the scoring of the field study data was done under more tightly controlled conditions (yielding a correlation of .77 between the two ratings on individual tasks), this evaluation was designed to provide an idea of how well the scoring model performs, given more varied distribution of candidates' scores for training and evaluation.

The slope and intercept parameters of the model were set to minimize the least squares error on the field study training data, and then the model was applied to the field study evaluation data, yielding the results in the final column of Table 12. Indeed, the greater variability of the human scores seems to make a large difference in the model's performance, as we achieve a correlation of .68 between the predicted score and the human-assigned score on this data set, for six items combined.

On the sm-eval and rec-train set, we also estimated the phi coefficient for the prediction of scores summed across six tasks. The phi coefficient for unrounded scores estimated by the CAC regression model was .93 and.85 for rounded scores. A similar analysis conducted on the human scores yielded a phi coefficient of .65 for single human ratings summed across six tasks, which resulted from the variability associated with both raters and tasks.

CART. In this particular problem, mixed priors (average of equal priors across score levels and the priors of the sm-train sample) with the Gini or Entropy splitting rules[3] gave comparable weighted kappas with the human scores on the cross-validation sample, among all the combinations. Then, the sm-eval sample cases were dropped down the best trees to obtain the classification rates.

The optimal tree using the mixed priors and the Gini splitting rule is presented in Figure 7. This tree shows visually the features that partitioned the sm-train cases into different score classes (terminal nodes) at certain splitting values. Conceptually, the splitting features and values defined the boundaries of different score classes. This is analogous to using responses that represent the lower and upper ends of a score class as range finders typically used in rater training.

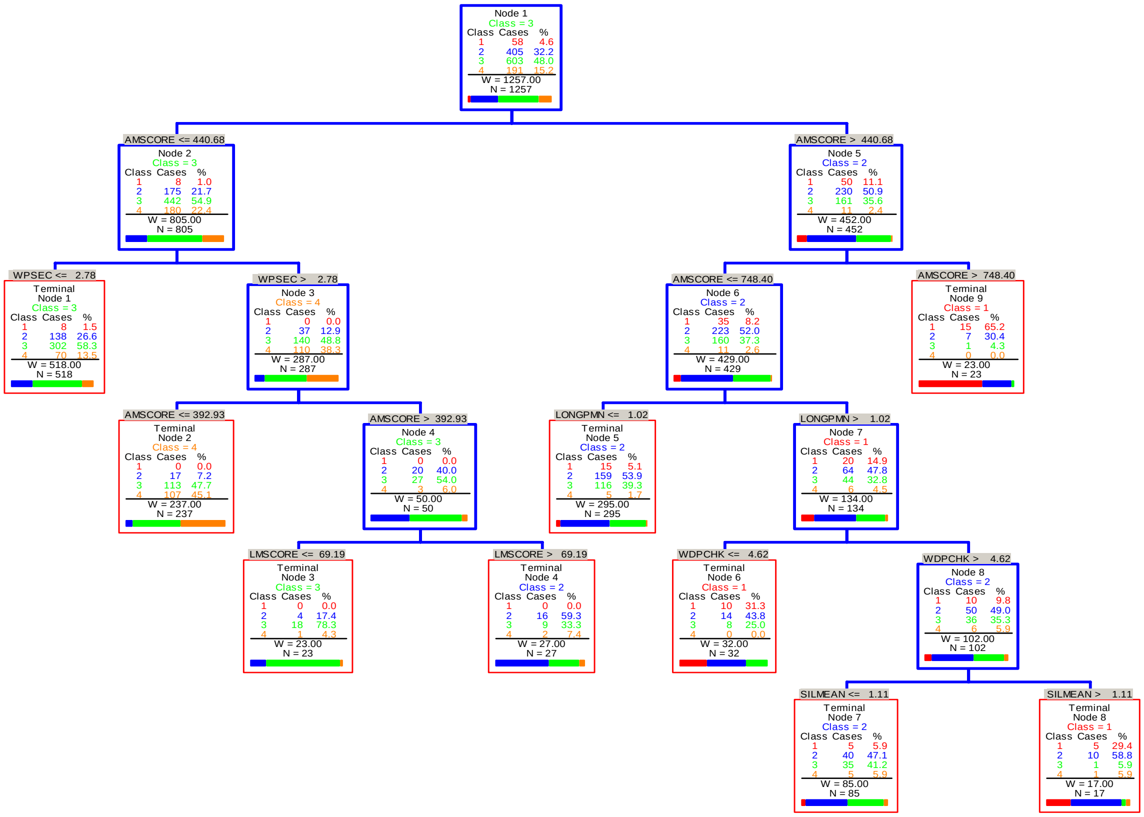


***Figure 7*. The optimal tree for classifying different score classes (mixed priors, Gini splitting rule).**

Two steps were followed by the CAC to review the substantive meaning of the tree structure. First, the splitting features and the relationships between the splitting features and the score classes were examined. Specifically, the CAC evaluated whether the decision rules that led to the classifications of students into different score classes (the terminal nodes) were consistent with their understanding of some typical profiles of students represented at each score level. The second step involved an examination of the splitting values at each decision point. Each splitting value was examined to ensure that the one selected by the tree algorithm corresponded empirically to the CAC's judgments. To facilitate the second step, borderline cases at each decision point (feature values close to the splitting value), along with some misclassified cases, were reviewed. As shown in Figure 7, five features were present in the tree: (a) amscore, (b) wpsec, (c) wdpchk, (d) silmean, and (e) lmscore. The first, amscore, was a pronunciation feature (Delivery). The second through the fourth were fluency features (Delivery). The last feature, lmscore, was a grammar feature (Language Use). All five features received very high ratings from the CAC in terms of their linkage to and representation of certain feature classes and dimensions in the rubric (see Table 8). Regarding the construct coverage of these features, the key pronunciation and fluency features were well represented, and the only grammatical feature was also present in the tree. However, tpsecutt, a very important vocabulary diversity variable, was not selected as a primary splitting variable in this empirically grown tree. It has to be noted that in a tree structure, one variable may obscure the significance of another, known as masking (Steinberg & Colla, 1997). The variable importance score indicates a variable's ability to mimic the functions of the primary splitters and to serve as replacements for primary splitters in the tree. This information can be used to select variables that are less expensive to collect or to enhance the substantive representation of the overall tree structure.

Based on the variable importance ranking information (Table 13), tpsecutt was ranked as the third most important variable. This suggests the possibility of using this variable to replace one or more splitters in the tree to improve the construct representation of the scoring solution.

Table 13

Variable Importance Ranking

<table>
<tr><th>Variable</th><th>Score</th><th></th></tr>
<tr><td>Amscore</td><td>100.00</td><td>||||||||||||||||||||||||||||||||||||||||||</td></tr>
<tr><td>Wpsec</td><td>59.07</td><td>||||||||||||||||||||||||</td></tr>
<tr><td>Tpsecutt</td><td>52.05</td><td>|||||||||||||||||||||</td></tr>
<tr><td>Longpwd</td><td>35.77</td><td>||||||||||||||</td></tr>
<tr><td>Wdpchk</td><td>34.44</td><td>||||||||||||||</td></tr>
<tr><td>Silmean</td><td>27.37</td><td>|||||||||||</td></tr>
<tr><td>Tpsec</td><td>18.35</td><td>|||||||</td></tr>
<tr><td>Silpwd</td><td>8.21</td><td>|||</td></tr>
<tr><td>Longpmn</td><td>7.17</td><td>||</td></tr>
<tr><td>Lmscore</td><td>5.87</td><td>||</td></tr>
<tr><td>Wdpchkmeandev</td><td>1.30</td><td></td></tr>
</table>

Now that we have examined the primary splitting variables, the decision rules that led to the terminal nodes are summarized in Table 14. These decision rules were considered by the CAC members to be reasonable. The different scoring rules for each score class were also deemed to be consistent with some of the typical profiles of students at a particular score level. However, a close examination of the cases in each terminal node would provide further confirmation that these profiles were typical. If they are determined to be nontypical upon further examinations, we can choose to remove the paths that led to the corresponding terminal nodes, because nontypical profiles are not very likely to generalize to new samples of candidates. Although we did not complete the second step, the work gave us some confidence that it was feasible to use expert judgments to examine the appropriateness of the splitting rules. The CAC

Table 14

Decision Rules for Different Score Classes

Rule	Terminal node	Performance characteristics
Score Class 1		
Rule 1	6	Poor pronunciation; pauses are on average long and chunks are very short (demonstrating very low automaticity)
Rule 2	8	Poor pronunciation; longer chunks but pauses are on average longer than those in responses in Terminal Node 6
Rule 3	9	Very poor pronunciation hinders the demonstration of other speaking skills
Score Class 2		
Rule 1	4	Good pronunciation but less strong grammar; speech rate too fast
Rule 2	5	Limited pronunciation skills; pauses are on average short
Rule 3	7	Limited pronunciation skills; pauses are on average longer, but word per chunk is more than that of responses in Terminal Node 6
Score Class 3		
Rule 1	1	Fair pronunciation and fast speech rate
Rule 2	3	Very fast speech rate; good pronunciation and good grammar (very fast speech rate does not pose much problem due to good pronunciation and grammar)
Score Class 4		
Rule 1	2	Very good pronunciation and very fast speech rate

members also noted that some other features, such as vocabulary sophistication and precision features and coherence and content relevance features, if available, may improve the accuracy in partitioning the cases into the right score classes. The optimal tree using the mixed priors and the Entropy splitting rule picked up two key variables, amscore and wpsec, as the primary splitters (Appendix D). This tree was nested in the tree examined in detail above, was less complex, but achieved almost the same level of accuracy. Although not explored in this study, it would be worthwhile to work with the CAC to compare these two tree structures and to take a hard look at which branches may generalize to new candidate samples, and thus should be kept, and which branches do not.

Table 15 shows the performance of the two optimal CART models. The reported statistics were the same as those reported for the multiple regression models, for comparison purposes. As discussed above, the correlation is the most important statistic we use to evaluate the model performance, along with RMSE, weighted kappa, mean and standard deviation of the predicted scores, and the exact-plus-adjacent agreement rates. For both CART models, the correlations with the human scores were .49 for individual tasks and .57 for scores summed across three tasks. The two models also showed almost identical performance as indicated by the weighted kappa.

For reasons mentioned above in the multiple regression results section, additional model evaluations were conducted with the TPO sm-eval and rec-train set and with the field study data set. When scores were aggregated across six tasks, we saw a correlation of .57 with the human scores. A CART trained on the field study training set was able to yield a correlation of .70 with human scores for sets of six tasks, indicating potential for improved performance with more variation in the scores (Table 16). As for the reliability of the scores, on the sm-eval and rec-train set, the phi coefficient for the scores summed across six tasks, estimated using the CART model (mixed priors and Gini splitting rule), was .90.

CAC Review. Table 17 shows the ratings of the substantive meaning of the CAC multiple regression model and the CART model (mixed priors, Gini) by four CAC members, using the rating form in Appendix C. On all three questions, the CAC members showed a preference for the CART model, especially on Questions 2 and 3. They thought that the CART model was a better representation of the relationships between automated features and human scores and that the way the model operates was more consistent with how expert raters decided on a score.

Table 15

Performance of Two Optimal CART Trees on TPO Scoring-Model Evaluation Set Data

<table>
<tr><th colspan="2">Model configuration</th><th colspan="4">Mixed priors,
Gini splitting rule</th><th colspan="4">Mixed priors,
Entropy splitting rule</th></tr>
<tr><td colspan="10">Single scores (<i>N</i> = 520)</td></tr>
<tr><td colspan="2">Quadratic-weighted κ</td><td colspan="4">.48</td><td colspan="4">.48</td></tr>
<tr><td colspan="2">Exact agreement</td><td colspan="4">54.62%</td><td colspan="4">55.77%</td></tr>
<tr><td colspan="2">Exact + adjacent agreement</td><td colspan="4">96.54%</td><td colspan="4">96.92%</td></tr>
<tr><td colspan="2">Mean (<i>SD</i>) of predicted score</td><td colspan="4">2.85(0.77)</td><td colspan="4">2.88(0.72)</td></tr>
<tr><td colspan="2">RMSE (rounded)</td><td colspan="4">.75</td><td colspan="4">.73</td></tr>
<tr><td colspan="2">Correlation (rounded)</td><td colspan="4">.49</td><td colspan="4">.49</td></tr>
<tr><td>Confusion matrix:</td><td></td><td><i>1</i></td><td><i>2</i></td><td><i>3</i></td><td><i>4</i></td><td><i>1</i></td><td><i>2</i></td><td><i>3</i></td><td><i>4</i></td></tr>
<tr><td><i>predicted</i> vs. human scores</td><td>1</td><td>5</td><td>10</td><td>3</td><td>0</td><td>3</td><td>12</td><td>3</td><td>0</td></tr>
<tr><td></td><td>2</td><td>8</td><td>82</td><td>58</td><td>11</td><td>1</td><td>86</td><td>61</td><td>11</td></tr>
<tr><td></td><td>3</td><td>2</td><td>60</td><td>165</td><td>62</td><td>0</td><td>58</td><td>169</td><td>62</td></tr>
<tr><td></td><td>4</td><td>0</td><td>2</td><td>20</td><td>32</td><td>0</td><td>2</td><td>20</td><td>32</td></tr>
<tr><td colspan="10">Pairs of scores (<i>N</i> = 232)</td></tr>
<tr><td colspan="2">Weighted κ</td><td colspan="4">.53</td><td colspan="4">.52</td></tr>
<tr><td colspan="2">Mean (<i>SD</i>) of predicted scores</td><td colspan="4">5.44(1.13)</td><td colspan="4">5.76(1.26)</td></tr>
<tr><td colspan="2">RMSE (rounded)</td><td colspan="4">1.22</td><td colspan="4">1.19</td></tr>
<tr><td colspan="2">Correlation (rounded)</td><td colspan="4">.55</td><td colspan="4">.54</td></tr>
<tr><td colspan="10">Triples of scores (<i>N</i> = 163)</td></tr>
<tr><td colspan="2">Weighted κ</td><td colspan="4">.55</td><td colspan="4">.54</td></tr>
<tr><td colspan="2">Mean (<i>SD</i>) of predicted scores</td><td colspan="4">8.67(1.95)</td><td colspan="4">8.78(1.81)</td></tr>
<tr><td colspan="2">RMSE (rounded)</td><td colspan="4">1.70</td><td colspan="4">1.65</td></tr>
<tr><td colspan="2">Correlation (rounded)</td><td colspan="4">.57</td><td colspan="4">.57</td></tr>
</table>

Note. CART = classification and regression tree, TPO = TOEFL Practice Online.

Table 16

CART (Mixed Priors, Gini) Model Performance on TPO Evaluation + Recognizer Training Set and on the TOEFL iBT Field Study Test Set

Model configuration		TPO evaluation + recognizer training set				Field study test set			
Single scores		*N* = 2,427				*N* = 1,752			
Quadratic-weighted κ		.43				.59			
Exact agreement		50.5%				50.6%			
Exact + adjacent agreement		94.8%				93.3%			
Mean (*SD*) of predicted score		2.88(.80)				2.47(0.92)			
RMSE (rounded)		.81				.73			
Correlation (rounded)		.44				.62			
Confusion matrix: *predicted* vs. human scores		*1*	*2*	*3*	*4*	*1*	*2*	*3*	*4*
	1	22	34	14	0	210	108	19	2
	2	67	305	273	64	101	275	162	15
	3	20	237	732	311	47	146	286	63
	4	2	26	153	167	6	31	166	115
Pairs of scores		*N* = 1,129				*N* = 854			
Weighted κ		.49				.66			
Mean (*SD*) of predicted scores		5.76 (1.42)				4.98 (1.67)			
RMSE (rounded)		1.34				1.44			
Correlation (rounded)		.50				.66			
Triples of scores		*N* = 757				*N* = 555			
Weighted κ		.53				.68			
Mean (*SD*) of predicted scores		8.71 (2.02)				7.53 (2.41)			
RMSE (rounded)		1.82				1.99			
Correlation (rounded)		.54				.68			
Sets of six scores		*N* = 308				*N* = 254			
Weighted κ		.55				.69			
Mean (*SD*) of predicted scores		17.44(3.80)				15.45(4.49)			
RMSE (rounded)		3.28				3.65			
Correlation (rounded)		.57				.70			

Note. CART = classification and regression tree, TPO = TOEFL Practice Online, iBT = Internet-based test.

Table 17

Evaluations of the Two Candidate Models by Four Content Advisory Committee Members

Question and judge	Multiple regression	CART
Question 1		
Judge 1	3.00	4.00
Judge 2	3.00	4.00
Judge 3	4.00	3.00
Judge 4	3.00	4.00
Avg.	3.25	3.75
Question 2		
Judge 1	3.00	4.00
Judge 2	2.00	5.00
Judge 3	4.00	4.00
Judge 4	3.00	4.00
Avg.	3.00	4.25
Question 3		
Judge 1	4.00	4.00
Judge 2	3.00	4.00
Judge 3	3.00	4.00
Judge 4	1.00	3.00
Avg.	2.75	3.75

Note. Judges 1 and 2 are responsible for training and monitoring the TOEFL iBT Speaking raters. Judgments on the classification and regression trees (CART) model were made before it could be modified for construct considerations, whereas the multiple regression model was an expert weight model endorsed by the Content Advisory Committee.

Although the models did not include any Topic Development features such as coherence, progression of ideas, and content relevance and did not cover the full spectrum of Language Use, they represented the Delivery features very well, especially fluency. Fluency is not a knowledge base on which speech production draws. Rather, fluency is related to multiple knowledge bases in speech production and manifests the degree of automaticity of deeper cognitive processes engaged during speech production. If a speaker does not have a large repertoire of lexical items, syntactic frames, sentence structures, or organization devices readily available or has difficulty pronouncing some words, his or her speed in lexical, grammatical, or phonological encoding is most likely to slow down. The speaker thus may sound less fluent. Therefore, fluency tends to be a key aspect of speech that indicates the *performance* level of a speaker.

Selecting the Final Scoring Model and the Score Reporting Method

For the TPO data, the correlation with human scores indicated that the CART model yielded a better performance than the multiple regression model in predicting individual task scores and scores summed across up to three tasks. However, the performances of the two models converged as the scores were summed across more tasks. As we will report only the total test scores averaged across the six tasks for SpeechRater v1.0, the statistics on the total test score were considered the most important in the model evaluation. As summarized in Table 18, for scores summed across six tasks, the correlations between the predicted scores and the human scores were the same (r = .57) for the two models, and the weighted kappa for the total score was higher for the CART model (.55) than for the multiple regression model (.51).

Table 18

Comparison of the Best Multiple Regression Model (CAC) and the Best CART Model (Mixed Priors, Gini Splitting Rule)

	Multiple regression model (CAC weights)		CART model (mixed priors, Gini splitting)	
Evaluation method for sets of six-item scores	TPO evaluation + recognizer training set	TOEFL iBT Field Study test set	TPO evaluation + recognizer training set	TOEFL iBT Field Study test set
Weighted κ	0.51	0.61	0.55	0.69
RMSE (rounded)	2.50	3.56	3.28	3.65
Correlation (rounded)	0.57	0.68	0.57	0.70

Note. CAC = Content Advisory Committee, CART = classification and regression tree, TPO = TOEFL Practice Online.

We obtained much higher correlations with the human scores for both models (.68 for the multiple regression model and .70 for the CART model) on the field study data. This suggests that the performance of the models, as indicated by the correlation with human scores, is likely to improve with data that show more variability in the scores and are more evenly distributed across the four score levels.

In terms of the generalizability of the scores, both the multiple regression and CART models produced scores that were highly generalizible across different tasks, the phi coefficient for scores summed across six tasks being over .90. Because automated scoring models could remove potential variability in the scores associated with raters, the dependability of the automated scores as indicated by the phi coefficient was actually much higher than that of the single human scores summed across six tasks (.65). When human raters were used, variability associated with both raters and tasks, and their interactions, contributed to the overall error variance in the scores. Therefore, if human rater agreement is poor, its adverse impact on the overall reliability of the scores may be severe.

In selecting the final scoring model, we considered the substantive meaning of the model, the mathematical and statistical principles underlying each model type, and the empirical results of each model. In addition to comparing the relative performances of the models, we evaluated whether the construct representation and the empirical performance of the models met the minimal requirements for use in a practice environment.

Despite the preference for the CART model by the CAC from the perspective of substantive meaning, both the multiple regression model and the CART model were judged to be adequate in representing the rubrics and in capturing the relationships between the automated features and the speaking construct for use in low-stakes practice settings, with ratings of 3 or above on a 5-point scale on Questions 1 and 2. The fact that the scoring models used only a subset of the features human raters use did not seem to have a serious adverse impact on the model performance, because the features included are key indicators of speaking performance, and different aspects of the speaking construct tend to be highly correlated (Xi & Mollaun, 2006).

Regarding the technical quality of the models, in general, multiple regression, as a parametric model that assumes a linear relationship between the features and human scores, is very efficient in its computation and can provide a good prediction when linear relationships are observed or approximated in the data. In contrast, CART does not assume that the human scores are particular functions of the features and allows the data structure to emerge from the data itself. Therefore, if the CART identifies multiple structures in the data, it may partition data into different regions and attempt to come up with a summary for the data structure in each region. This method may work well if, indeed, strong nonlinear relationships are present in the data. As for sample size requirement, multiple regression requires a much smaller sample to produce a stable solution than CART, because in CART the relevant data become less for each region, requiring a large sample to yield a stable solution.

With regard to this particular problem, with some of the nonnormal features transformed and outliers preprocessed to improve the normality of the features and the linear relationships between the features and the human scores, the multiple regression solution yielded fairly similar results in predicting the *total* test scores as the CART model that used raw feature values. This level of agreement was acceptable for a low-stakes application such as the TPO, given that we obtained much higher correlations on the field study data, which were more varied and more evenly distributed across the four score levels.

Although the CART model was superior in predicting the individual task scores and scores averaged across up to three tasks, we report only the total test scores averaged across six tasks in the SpeechRater v1.0. To this end, the performance of the multiple regression model appeared to be adequate.

With regard to prediction bias, the multiple regression model produced lower bias than the CART model in the predicted total test scores, as shown in the means and the associated RMSE estimates. The multiple regression model was able to reproduce the mean of the total human test scores better than the CART model, and the RMSE estimate of the multiple regression model was smaller as well.

Whereas the CART model was preferred by the CAC from a substantive perspective, the multiple regression model was favored by the TAC for its stability and parsimony. A multiple regression model that uses fixed weights based on expert judgments is also more flexible, compared to a CART model, if we need to modify the model to match the parameters of shifting TPO user populations (i.e., changes in means and standard deviations of scores). With the TOEFL iBT still rolling out in more countries and regions, some shifts in the TPO user populations are expected to occur. Therefore, we need to monitor closely the population shifts and to update the scoring model to match the new population parameters if necessary. An expert-weight multiple regression model would clearly be advantageous compared to CART for such efforts, as it would only involve rescaling to the mean and standard deviation of the new population. As for CART, priors that match the score distributions of the new population can be used in case of a population shift, but this is likely to lead to some changes in the model structure (i.e., splitting features and splitting values), which would require extensive technical and substantive reviews.

Another consideration in choosing the final scoring model is that multiple regression is more accessible to the general measurement and language testing audience. Although CART has some characteristics that are compatible with the speaking construct and that are consistent with the way task-level speaking scores are assigned by human raters, as a more novel methodology, it may take some time for it to make its way into the mainstream measurement and testing literature. At the initial stage of launching an automated scoring system, it is especially important to adopt a scoring methodology that is both effective and easy to explain to the general public. Given the above considerations, a decision was made to use the multiple regression model for the SpeechRater v1.0.

7. Development and Validation of a Filtering Approach

Importance and Role of a Filter

In our earlier discussion, we have focused on the task of assigning a score between 1 and 4 to an individual student response. However, the full task to be addressed is somewhat more complicated. To build a system that could score all incoming audio responses, we must first determine whether the audio stream represents a scorable response (0 or TD vs. 1–4). The model that makes this determination is called the *filtering model*.

The filtering model is applied to a response before it is sent to the regular scoring model. Its purpose is to determine whether a response constitutes a legitimate attempt to respond to a task and ought to receive a real score (1–4), or whether it is seriously anomalous in some way and ought to receive a 0 or a TD.

The filtering model is an integral part of the whole scoring method, and the results associated with it provide evidence for the Evaluation inference that the SpeechRater scores are accurate in representing the quality of a candidate’s performance on the practice test. The development and validation of this model is presented as a separate section for the sake of clarity.

Approach to Filter Development

We found that it was simple to build a model that discriminates between scores of 1–4 on the one hand and TD or 0 on the other. However, discriminating between TDs and 0s was much more difficult. The rubric used for the TOEFL iBT Speaking test defines the scores of 0 and TD in such a way that they are quite difficult to distinguish reliably, even for trained human scorers. Briefly stated, a score of 0 is assigned if the speaker was unwilling or unable to respond or made no attempt to answer the question, whereas a TD is assigned under a number of special conditions, such as if the response is too loud, too quiet, contains noise or feedback, or contains complete silence. The difficulty in distinguishing between the two score classes arises because by far the largest class of anomalous responses includes those that consist almost completely of silence. In such cases, the distinction between the two classes boils down to whether the scorer hears evidence of the candidate's presence, such as breathing. If the candidate is thought to be at the microphone, the response is scored as a 0. Otherwise the candidate receives a TD.

In a low-stakes application like the TPO, the penalty for not approaching each task in the way intended by the assessment design is very low; the only disincentive is that this would be a waste of money and an opportunity for practice on the task. Consequently, the chance of our model seeing anomalous responses is higher than in the TOEFL iBT, where few candidates are likely to risk getting a low score by failing to respond to a task.

On that subset of our TPO data that both human raters scored as either a TD or a 0, the agreement between raters was very low (46.5%; κ = .094). It should be mentioned that the two sets of raters assigned different proportions of 0s and TDs, so it is possible that better agreement could be achieved if pains were taken to ensure consistent training of both groups of raters on this task.

The difficulty in discrimination between 0s and TDs was apparent in our attempt to build the filtering model as well. None of the features we looked at was of any use in discriminating between responses assigned the scores 0 or TD.[4] The features we considered included all of the features investigated for use in the scoring model as well as an additional set of rough prosodic features, consisting of the moments of utterance pitch and power. (In particular, a measurement of the average acoustic power of an utterance was considered a potentially promising way to determine whether a response contained any speech at all.)

In operational TOEFL iBT testing, whereas 0s are taken into account in computing the total scores, candidates are offered another chance to take the test if two or more TDs occur out of the six responses. However, given the purpose of the TPO, it is less crucial to make this distinction between 0 and TD than in operational testing, as we can offer candidates a second chance to respond in both cases. Therefore, we decided to construct a model to classify responses as either *scorable* or *anomalous*. Scorable responses are passed on to the regular scoring model for processing, whereas anomalous responses are treated as TDs in the practice environment, meaning that candidates are given an opportunity to rerecord their response.

Design of the Study Evaluating the Filter

In developing and evaluating the filtering model, we used the portion of the TPO data that consist of responses scored either 0 or TD by human raters. The data were subdivided into a training portion, *filter-train*, and a test portion, *filter-test*. The sm-train (described in Section 6) and filter-train sets were combined to produce the complete set of training data for this task (and similarly with the evaluation set), and scores were collapsed to pose the problem as a binary classification task. Responses with scores of 1–4 were labeled as scorable, whereas responses with scores TD or 0 were labeled anomalous. This left us with 1,595 responses to use for training, of which 338 were anomalous, and 660 responses for model evaluation, of which 140 were anomalous. The summary statistics of the TPO data sets used in developing and evaluating the filtering model are presented in Table 19.

Table 19

Summary Statistics of the TPO Data Sets Used for Filtering Model Development and Evaluation

Data set	No. responses	No. speakers	No. topics	Avg. score	*SD* of score	Score distribution					
						TD	0	1	2	3	4
sm-train	1,257	263	15	2.74	0.77	0	0	58	405	603	191
sm-eval	520	120	9	2.73	0.69	0	0	18	159	289	54
filter-train	338	93	15	N/A	N/A	220	118	0	0	0	0
filter-test	140	41	9	N/A	N/A	99	41	0	0	0	0

Note. TPO = TOEFL Practice Online; TD = technical difficulty; sm-train = scoring-model training, sm-eval = scoring-model evaluation; filter-train = training portion of data; filter-test = test portion.

Table 20, which presents the confusion matrix for human ratings of the responses in this combined data set, shows an overall exact agreement rate of 55.0% on single items, but this number is not very revealing. For a more informative assessment of the agreement levels between human raters, we need to break down the scoring process in the same way SpeechRater does. First, we consider the division of responses into the scorable class of responses, which are labeled 1–4, versus the anomalous class of responses, which are scored as TD or 0. Then, we determine the level of human agreement within each of these classes.

If we conflate the score classes TD and 0 on the one hand and the scores 1–4 on the other, human agreement on this simpler distinction is quite good. The two sets of raters agreed with one another in classifying responses as scorable or anomalous 98.8% of the time (κ = .940). Agreement in classifying anomalous responses as either 0 or TD (the upper left four cells of Table 20) was much lower. The exact agreement was only 46.5% (κ = .094).

Table 20

Confusion Matrix Showing Single-Item Agreement Between Two Sets of Human Raters on Full Set of TPO Data

		Rater 2						
		TD	0	1	2	3	4	Total
Rater 1	TD	75	20	2	3	0	0	100
	0	225	138	3	0	0	0	366
	1	4	2	47	21	3	0	77
	2	9	2	66	582	347	17	1,023
	3	13	6	10	475	1,148	190	1,842
	4	4	3	1	52	430	340	830
	Total	330	171	129	1,133	1,928	547	4,238

Note. TD = Technical difficulty score, TPO = Test of English as a Foreign Language Practice Online.

Analysis

Fourteen speech features indicating the power and pitch characteristics of a response, the number of types of words in a response, and the recognizer's overall confidence about its recognition results were used as the candidate features for building the filtering model (cf. Table 21). The first model that we considered simply used the training set to establish an optimal cutoff on a single feature. Although we did not expect this model to perform as well as the more complex model introduced below, we still might have used this single-feature model operationally if the performance difference was minor, because of its simplicity. The feature that gave us the best discrimination in our data set was types (the number of distinct words recognized). The trained model labeled all responses with fewer than 33 distinct word types (which was the optimal value for classifying the training set) as anomalous.

Table 21

Candidate Features for the Development of the Filtering Model

Feature name	Feature description
1. Powmean	Mean global power
2. Powmeandev	Mean deviation of power
3. Powvar	Variance of power
4. Powstddev	Standard deviation of power
5. Powmin	Power global minimum
6. Powmax	Power global maximum
7. Powdelta	Power—difference between max and min (powmax – powmin)
8. Pitmeandevnorm	Pitch mean deviation normalized by mean pitch
9. Pitminnorm	Minimum pitch normalized by mean pitch
10. Pitmaxnorm	Maximum pitch normalized by mean pitch
11. Pitdeltanorm	Difference of maximum and minimum pitch normalized by mean pitch
12. Types	Number of word types; i.e., unique word forms[a]
13. Confavg	Average overall confidence scores in a sample[b]
14. Conftimeavg	Average of time-weighted confidence scores

[a] Word forms are not stemmed; e.g., *pet* and *pets* are two types. [b] Every word has one confidence score associated with it.

The other model we considered used four features: (a) types (number of distinct words recognized), (b) confavg (average recognizer confidence score), (c) powmean (average power of speech signal), and (d) powmeandev (mean absolute deviation of speech signal power). Intuitively, these features ensure that the likelihood of a response being labeled anomalous will increase as (a) it contains fewer recognizable words, (b) the recognizer is less sure of its hypothesis, (c) it contains less sound, and (d) the variability of the sound level is lower.

We used a classification by regression procedure to build this second model (Frank, Wang, Inglis, Holmes, & Witten, 1998). This means that we treated the class membership as a continuous variable (0 for anomalous responses and 1 for scorable ones) and then found the optimal cutoff on the regression value obtained. Feature values were standardized before model training for maximal interpretability of the weights.

We used 10-fold cross-validation within the training sample to experiment with features and model configurations. Once we had converged on these two final models, we evaluated both on the test set.

Results

The results of these two candidate filtering models are provided in Table 22. In Table 22, *overall accuracy* is the proportion of all responses which each model classifies correctly. *Precision* is the proportion of the responses classified as anomalous that are indeed anomalous. *Recall* is the proportion of anomalous responses that our model correctly finds. *False positive rate* is the proportion of scorable responses that are misclassified as being anomalous. This final number, the false positive rate, is of greatest importance for our application. Although we do not want anomalous responses to receive a score, the cost of incorrectly filtering out a good response is much greater.

Table 22

Filtering Model Results in Percentages

Model	Overall accuracy	Precision	Recall	False positive rate
		Using 10-fold cross-validation on training set		
One feature	97.8	96.9	92.6	0.8
Four features	98.3	99.1	92.9	0.2
		Test set		
One feature	98.8	98.5	95.7	0.4
Four features	99.2	100.0	96.4	0.0

Table 22 shows that the test set is somewhat easier than the training set and that our models generalize very well. Both models allow us to find over 90% of the anomalous responses, while keeping the false positive rate well under 1%. In our production application, we have chosen to use the four-feature model, because of its higher agreement with human raters, and because it is less reliant on a single source of information (the number of distinct words recognized). The feature weights in the multiple regression model were .56 for types, .09 for confavg, .14 for powmean, and .22 for powmeandev, so the types feature is clearly still the most important factor in this model.

8. Development of the User Interface

Having developed a scoring model that can predict scores based on the selected scoring features, we are left to address the user interface, which will convey the score information to users, as well as other relevant advisories that facilitate the interpretation and use of SpeechRater scores. In this section, we discuss the construction of the prediction intervals that convey the degree of uncertainty around the SpeechRater scores. We also describe the development of the score report and user advisories that communicate the limitations of SpeechRater v1.0 and stress the low-stakes use of the scores.

This section presents evidence that can be fed into the Utilization inference that pertains to the relevance, sufficiency, and usefulness of the SpeechRater scores for the intended application. We also recognize that the evidence directly supporting the preceding inferences is relevant to the Utilization inference as well.

Prediction Intervals

In addition to our score prediction provided by the multiple regression scoring model, we provide an indication of the expected amount by which this predicted score might differ from the score a human would assign to a response, or a set of responses. In short, we provide a *prediction interval*.

One way to do this would have been to provide a symmetric band, centered around the predicted score, based on the standard error of measurement. However, because the error for the extreme score levels is usually smaller than that for the middle score ranges, we chose instead to provide an interval based on a cumulative logit model, which estimates the likelihood of a predicted score corresponding to a human score at each point in the score scale. Given these probabilities associated with each point in the score scale, we then could find a central region with a probability mass above a certain value and report it to the examinee as the interval within which a human's rating would be expected to fall with a certain probability.

A 90% prediction interval was chosen, because it contains enough of the probability mass that the examinees could be fairly certain that their scores would fall within that range, while allowing for relatively simple explanation (e.g., "9 times out of 10, a trained human rater would score your response within this range").

Figure 8 shows the 90% prediction intervals associated with each predicted score, for a full set of six tasks. Because we do not report any score but the total speaking section score on the TPO assessment, we did not need to create prediction intervals for individual task scores.

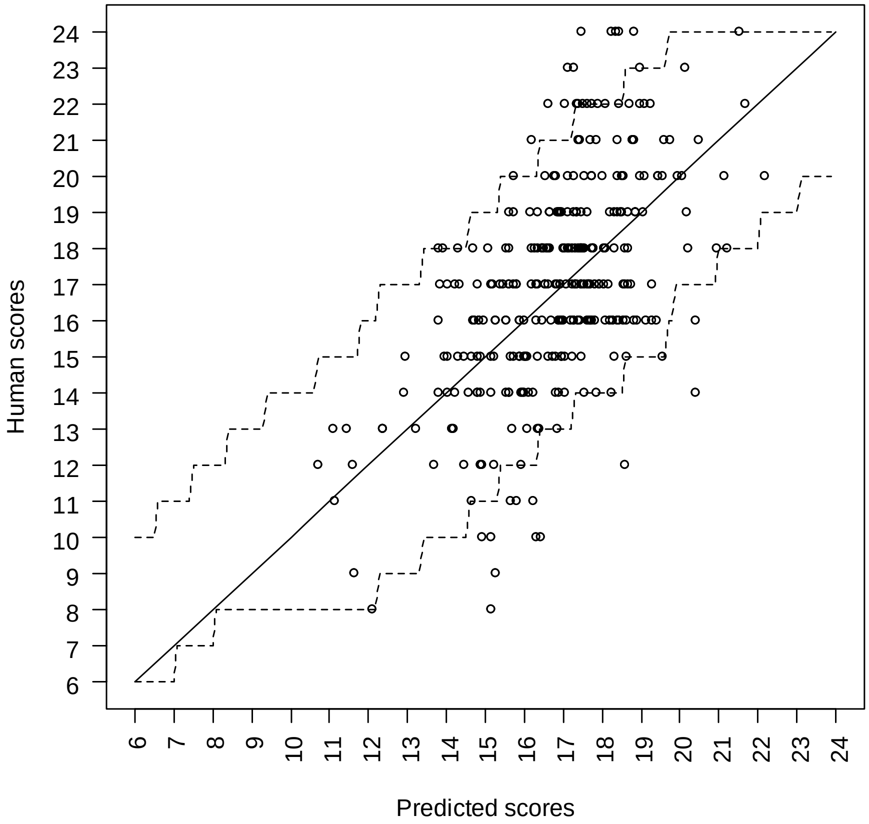


***Figure 8.* Prediction intervals for the full range of predicted speaking section scores.**

In Figure 8, the prediction interval for a given score predicted by SpeechRater v1.0 is to be read vertically, from the lower dotted line to the upper dotted line. So, to find the interval for a predicted score of 14, we read across the x-axis until we get to the 14 tick-mark and then follow a vertical line straight up through the dotted lines at 10 and 18. Thus, for a predicted score of 14, the 90% prediction interval is (10, 18). The 90% prediction interval averages about 8 score points on the 0–24 raw total score scale. (It is a bit wider when this score is converted to the 0–30 scale on which TOEFL iBT section scores are reported.)

The cumulative logit model was trained on the combination of the sm-eval set and the rec-train set, because the former data set did not contain enough candidates with complete sets of six task scores. It is less than optimal to use data on which the speech recognizer was trained in further model building, but realistically we do not believe that there is any significant risk of any sort of overtraining that would corrupt the calculation of prediction intervals. It was necessary to combine these two data sets in order to have enough examinees with six completed items, so that we could carry out the analysis. Figure 8 shows the data from this combined set, plotting each correspondence between human score and predicted score for a set of six tasks as a point on the graph. The empty region in the lower left of the graph shows that our data contained very few examinees with aggregate scores below 10.

Score Report and User Advisories

The score report is a critical piece of an automated scoring system because it communicates to the users how the scores should be interpreted and used. The score report of the TOEFL iBT Speaking Practice test (Appendix E) reports both the total scaled score and the possible scores that one may receive if the test were scored by trained human raters. No individual task-level scores are reported. Because task-level scores were much less reliable, reporting only the total score summed across six tasks ensured that the reliability of the reported scores was at an acceptable level. In addition, the automated score had a much higher correlation with the human scores at the test level than at the task level. Therefore, the test-level scores were also more valid if human scores were used as the criterion. The range of human rater scores for a particular predicted automated score is included in the score report to communicate the uncertainty around the predicted total scores.

A direct link to this score report, "How Your Practice Test Was Scored" (Appendix F), provided three key pieces of information:

1. First, the SpeechRater v1.0 uses only a subset of the criteria used by human raters to score the TOEFL iBT Speaking test.

2. The SpeechRater scores are just estimates of candidates' performances on the TOEFL Speaking Practice test.
3. The score range should be interpreted as the range human scores would fall within 90% of the time.

Another document that is linked to the score report, "Frequently Asked Questions About TOEFL Practice Online Automated Scoring for Speaking" (FAQs; Appendix G), is intended to elaborate on the key points covered in "How Your Practice Test Was Scored" as well as communicate some additional messages. For example, this document cautions against using the TOEFL Speaking Practice Test scores to predict scores on TOEFL iBT Speaking test, which is taken under regular testing conditions, because no study has been completed yet to compare candidates' performances on these two tests. These three documents convey to the users important information about the capabilities and limitations of the SpeechRater v1.0 and how these would impact the interpretations and uses of the SpeechRater scores.

9. Discussion and Conclusion

Evaluation of the Strength of the Overall Validity Argument

We have attempted to develop different pieces of evidence required to support the interpretive argument about SpeechRater v1.0 as stated in Section 4. We now have come to the final step of integrating the evidence and evaluating the plausibility of the interpretative argument in the context of a validity argument. It is useful to revisit the claim we would like to support for SpeechRater v1.0.

The SpeechRater v1.0 score is a prediction of the score on the TOEFL iBT Speaking Practice test a test taker would have obtained from trained human raters. The entire practice experience can help familiarize test takers with the content and format of the TOEFL iBT Speaking test so that they can better prepare for it. This score can be used by the test takers to help them self-evaluate their readiness to take the TOEFL iBT Speaking test.

This claim clearly specifies the intended low-stakes use of the TOEFL iBT Speaking Practice test and the score that SpeechRater v1.0 produces. Although this claim states what the SpeechRater v1.0 intends to do, it also conveys, although not explicitly, what it does not do. First, it does not intend to predict a candidate's potential performance on the TOEFL iBT Speaking test, which is taken under operational testing conditions. The motivation and anxiety levels of the candidates may be different when taking the official test versus the practice test. When taking the real test, candidates may be more motivated but more nervous. In addition, candidates can make several attempts on each task in the practice test whereas they are allowed only one attempt on each task in the official test. When taking the practice test, candidates also could choose to use more time to plan a response before starting to record it, but this option is not available for the official test.

Second, SpeechRater v1.0 does not intend to *explain* why a candidate receives a certain score. More specifically, the scoring model of the SpeechRater v1.0 does not mimic exactly how a human rater would have scored a test. It only intends to use meaningful speech features that indicate different aspects of candidates' speaking performance to *predict* the score of a human rater.

Given that this initial version of SpeechRater focuses on providing prediction of human scores at a level acceptable for low-stakes decisions in practice environments, three of the five inferences particularly need to be backed by relevant empirical or judgmental evidence: (a) Evaluation, (b) Generalization, and (c) Utilization. The Evaluation inference pertains to the accuracy of the automated scores; the Generalization inference concerns the stability of the scoring model and the generalizability of the scores across different tasks; and the Utilization inference is related to the sufficiency, relevance, and usefulness of the score and other related information provided to candidates for making self-evaluations of their speaking performance. Although the Extrapolation and the Explanation inferences are important, adding meaning and value to the SpeechRater scores to support the subsequent Utilization inference, it is less critical for them to be fully supported for this version of SpeechRater.

Table 23 displays summaries of the evidence collected that is pertinent to the rebuttals that may weaken each of the inferences supporting the intended use of the scores. With regard to the Evaluation inference, the predicted scores by SpeechRater v1.0 were able to achieve an agreement with the human scores at a level considered to be reasonably adequate for a low-stakes application. The appropriate model training and evaluation procedures also ensured the generalizability of the scoring model to new tasks and new samples of candidates. In addition, the predicted scores were shown to be highly generalizable across tasks, lending further support to the Generalization inference. We do not, however, have any evidence to show whether this decrease in the variability of SpeechRater scores associated with tasks in comparison to the human scores was desired. Because we have not collected the candidates' scores on an

Table 23

Evidence Collected to Discount the Rebuttals That Weaken Each Inference

Rebuttals	Counterevidence
Evaluation: Automated scoring results in scores that accurately represent the quality of the performance on the practice test.	
1. The scoring algorithm under- or misrepresents the construct or introduces construct irrelevance so that the resulting scores are not accurate.	The filtering model effectively separated scorable (scores 1–4) and nonscorable responses (0 or TD). We were not able to build a model that distinguished 0s from TDs well. However, this 0 vs. TD distinction is less critical for the practice test. The correlation between the automated and the human scores summed across six tasks was moderate (.57) but may improve with data that demonstrate more varied proficiency levels. The weighted kappa between the human and the automated scores summed across six tasks (.51) was only moderately high but may be higher with data that demonstrate more varied proficiency levels.
Generalization: The scoring model can generalize to new tasks and samples of candidates, and the automated scores are generalizable over tasks.	
1. The scoring model is built from insufficient or unrepresentative samples.	The model training sample was large (*N* = 1,257) and was representative of the candidates who have taken the TOEFL iBT Practice Speaking test. Although we might see population shifts after the TOEFL iBT rolls out in more countries and regions, concrete plans have been drawn up to monitor the population changes in relation to the phased roll-out plan and modify the scoring model accordingly.
2. The scoring model does not generalize to new tasks or independent candidate samples.	The scoring model was cross-validated on an independent testing sample (*N* = 520) that did not have the same tasks and candidates as in the training sample in tasks and candidates, supporting the generalizability of the scoring model to new tasks and to new samples of candidates.
3. The automated scores do not generalize across tasks.	The phi coefficient for the automated scores for six tasks was very high, indicating that the automated scores were generalizable across different tasks. That is to say, a candidate who takes a new set of tasks is likely to receive a similar automated score on the test.

(Table continues)

Table 23 (continued)

Rebuttals	Counterevidence
Extrapolation: The automated scores reflect the quality of performance on relevant real-world speaking tasks in an academic environment.	
1. Candidates' automated scores are not related to their levels of performance on real-world speaking tasks in an academic environment.	No evidence has been collected to discount this rebuttal.
Explanation: The automated scoring model captures aspects of performance that reflect the underlying speaking abilities used in an academic setting.	
1. The automated scores are not adequate in *explaining* examinee performance in the domain.	Based on Brown, Iwashita, & McNamara (2005), the rubrics for the TOEFL iBT Speaking test were reflective of what teachers of English as a second language and applied linguists thought were important in evaluating candidates' speaking performance in an academic environment. However, the features used in the automated scoring model were only a subset of the criteria used by the human raters, reducing the model's power in explaining candidates' performance on real-world speaking tasks.
2. The speech features used in scoring models are not well linked to the rubric, introducing construct irrelevance.	Based on the content specialists' ratings, the speech features were reasonably well linked to certain aspects of the rubric.
3. The speech features do not cover the key criteria defined in the rubric very well, resulting in construct underrepresentation.	Based on the content specialists' ratings, the speech features combined covered the TOEFL iBT rubric only moderately well.
4. The speech features are not combined in a meaningful way to produce scores.	The weights used to combine the features values to produce the automated scores were based on expert judgments and endorsed by the content specialists.
5. The scoring model disproportionately captures aspects of the rubric that generalize across tasks, reducing task specificity in an undesirable way so that the constructs are underrepresented.	No evidence has been collected to discount this rebuttal.

(Table continues)

Table 23 (continued)

Rebuttals	Counterevidence
Utilization: The automated test scores and other related information provided to candidates are *relevant, useful,* and *sufficient* for them to make intended decisions and promote positive effects on teaching and learning.	
1. The predicted scores and other information communicated to the candidates do not provide relevant, useful, and sufficient information for them to gauge their readiness to take the TOEFL iBT Speaking test.	The 90% prediction intervals were fairly wide but were acceptable for candidates to use to get a sense of their readiness to take the TOEFL iBT Speaking test. The automated score feedback provides an estimate of a candidate's scores on the TOEFL iBT Speaking Practice test that human scorers would have assigned. Therefore, they may be able to evaluate their level of speaking proficiency and make efforts to improve it if their scores are low.
2. The automated scores negatively impact users' perceptions of the assessment and the way they interpret and use the scores as intended.	No evidence has been collected to discount this rebuttal.
3. The potential negative consequences of SpeechRater v1.0 are not anticipated and minimized.	The distinction between automated scoring for the TOEFL iBT Speaking Practice test and human scoring for the TOEFL iBT Speaking test is explicitly defined in the FAQs so that adverse impact on the credibility of the scoring of the TOEFL iBT Speaking test is minimized. The limitations of SpeechRater v1.0 are clearly communicated to the candidates through the two documents linked to the score report to caution against the use of the scores for important decisions.
4. The automated scoring system does not promote positive washback effects on English language teaching and learning.	No evidence has been collected to discount this rebuttal.

Note. iBT = Internet-based test; TD = technical difficulty.

appropriate criterion measure, we are not able to make that judgment. Thus, we cannot conclude how this would impact the Explanation inference.

The scoring model used some key speech features considered to be meaningful by the content specialists and combined these features to produce scores in a way that was consistent with how the content specialists thought they should contribute to the human score. However, we have to note that the features included in the scoring model were only a subset of the criteria that human raters use, reducing its power in *explaining* all of the key speaking skills underlying a candidate's performance.

With regard to the Utilization inference, the evidence discussed above for the Evaluation inference provides some support for the acceptable level of accuracy in the model's prediction of human scores. Building on the evidence, we need to demonstrate further whether the predicted scores allow the candidates to evaluate their own readiness to take the official test, which is how the SpeechRater scores are intended to be used. As discussed earlier, SpeechRater is not intended to predict candidates' scores on the official test, given possible differences in the motivation level and test-taking conditions. However, a candidate may be able to self-evaluate his or her readiness for the official test, knowing the conditions under which he or she has taken the practice test. A candidate could choose to take the practice test under the timed mode and make his or her best effort to respond to each task as if he or she were taking the official test. Only when taken under these circumstances would a candidate be able to assess his or her own readiness to take the official test.

Error in the SpeechRater predicted scores on the practice test may impact the decision-making processes of candidates who take the practice test to gauge their readiness to take the test. Receiving lower scores than deserved may discourage a candidate who is ready from taking the test, whereas getting higher scores than deserved may encourage a candidate who is not ready to take the test. Most likely, candidates may decide whether to take the official test after comparing their scores on the practice test to the admissions standards of the programs to which they are interested in applying.

Based on the minimal score requirements on the TOEFL iBT Speaking test set by different schools for admitting international students into U.S. universities (ETS, n.d.), it appears that most schools select cut scores in the range of 20–23 on the scaled score scale (16–18 on the raw score scale) for the TOEFL iBT Speaking test. These cut scores translate into an average raw task score of 2.7–3.0 on a scale of 0–4 points. This suggests that for the TOEFL iBT Speaking Practice test, it is especially important for us to provide a good prediction for candidates whose raw total scores are in the range of 13–20 (i.e., those whose scores are around or just below the raw total cut scores). Candidates scoring in this range are likely to be "on the fence" regarding their confidence in taking the official test and would need reasonably accurate score information to evaluate their own readiness for taking the official test.

As shown in Figure 8, candidates' scores were overpredicted at the very low end (raw score total in the range of 6–12), but the possible human scores would still be under 16, which are lower than the typical cut scores for admissions. Therefore, even if these low-scoring candidates received somewhat higher scores than warranted, generally they would not receive enough encouragement to think they were ready to take the official test. In contrast to the scores at the lower end, the scores at the very high end (raw score total in the range of 21–24) were underpredicted. However, for a practice test preparing candidates to take the TOEFL iBT Speaking test, the prediction error at the high end is less serious, thus not jeopardizing the intended use of the practice product. The tendencies of the middle range scores being over- or underpredicted by the SpeechRater were relatively low, compared to the scores at the higher or lower end. For predicted raw total scores in the range of 13–20, the 90% interval spans about 8 score points on the 0–24 scale (Figure 8) and includes the range of typical admissions cut scores (16–18 on the raw score scale). The 90% interval around the SpeechRater scores was fairly wide. However, given that the score information may only be used by the candidates to determine their readiness for the official test or to help them practice their speaking skills, it was deemed to be acceptable for this purpose. Furthermore, the limitations of the SpeechRater v1.0 were explicitly conveyed to the users, including its underrepresentation of the speaking construct and its imperfect prediction of the human judgments on the practice test. Being forthright with the users about the limitations of the SpeechRater v1.0 could reduce the chances that the scores are misused for high-stakes decisions and mitigate potential negative consequences. In sum, the evidence collected in this project, when taken together, provides fairly adequate, if not strong, backing for the claim we wish to make for the SpeechRater v1.0.

Decision to Release the Model to Operational Use on a Conditional Basis

Ultimately, the construct representation in the CAC regression model was sufficiently broad to justify its use in low-stakes applications. While higher order parts of the speaking construct (such as grammar and topic development) are missing, or imperfectly modeled, more basic aspects of the construct (such as pronunciation and fluency) are richly represented. In addition, these different parts of the speaking construct tend to be highly correlated, so that the absence of higher order factors is not as detrimental to the model's agreement with human raters as it otherwise might be.

The model's agreement with human raters was not as high as we would have liked but is still suitable for use in low-stakes applications. The correlation of the six-item aggregate score with human raters is the best indication of model quality for this data, since this aggregated score is the only number we will report. Further, the correlation of .57 reported above is acceptable, given the low human agreement on this TPO scoring task and the fact that we obtain a much higher correlation of .68 on data with more variability in the scores, such as the field study data. Furthermore, the phi coefficient of the CAC regression model predicted scores was quite high for the six tasks, supporting the high degree of generalizability of scores across tasks.

Given the limited construct representation and modest prediction accuracy of the CAC scoring model, recommendations were made to the TOEFL program to release the model for use in the TPO with the following conditions:

1. Prediction intervals must be reported to indicate the error around the automated scores.
2. Limitations of this version of SpeechRater must be communicated.
3. Distinction between the scoring for the TPO and for the TOEFL iBT must be stressed.
4. The low-stakes practice use of the scores must be emphasized.

Conclusion

This study reports the development of the SpeechRater v1.0 system and its validation for low-stakes practice purposes using an argument-based approach. The processes we followed to build this system represent a principled approach to maximizing two essential qualities of an automated scoring system: substantively meaningful and technically sound. The argument-based approach to validation provided a mechanism for us to articulate the strengths and weaknesses in the validity argument for SpeechRater v1.0 and put forward a transparent argument for using it for a low-stakes practice environment. An inherent advantage of this approach is that it allowed us to identify critical gaps in our existing research for SpeechRater v1.0 and allocate resources to address these gaps in our future research. Specifically, the areas of research to pursue include improving the prediction accuracy for the whole test-taking population and for test takers with different native language backgrounds and expanding the construct coverage of the scoring model. Furthermore, we need to explore alternative criterion measures other than human scores to validate the automated scores. Other critical areas of investigation include users' perceptions of and interactions with this system and the impact of users' perceptions on their decision making based on the scores.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2.0. *Journal of Technology, Learning, and Assessment, 4*(3), 1-34.

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

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1-34.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford, England: Oxford University Press.

Bejar, I. I. (1991). A methodology for scoring open-ended architectural design problems. *Journal of Applied Psychology*, *76*(4), 522-532.

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice*, *17*(4), 9-17.

Bennett, R. E., Sebrechts, E., & Marc, M. (1994). *The accuracy of automatic qualitative analyses of constructed-response solutions to algebra word problems* (ETS Research Rep. No. RR-94-04). Princeton, NJ: ETS.

Bernstein, J. (1999). *PhonePass testing: Structure and construct*. Menlo Park, CA: Ordinate.

Braun, H. I., Bennett, R. E., Frye, D., & Soloway, E. (1990). Scoring constructed responses using expert systems. *Journal of Educational Measurement*, *27*(2), 93-108.

Brieman, L., Jerome F., Olshen, R., & Stone, C. (1984). *Classification and regression trees.* Pacific Grove, CA: Wadsworth.

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 No. 29 MS-29). Princeton, NJ: ETS.

Burstein, J., Kukich, K., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., et al. (1998). *Computer analysis of essay content for automated score prediction: A prototype automated scoring system for GMAT analytical writing assessment* (ETS Research Rep. No. RR-98-15). Princeton, NJ: ETS.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 8-24.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. Mahwah, NJ: Lawrence Erlbaum.

Chodorow, M., & Burstein, J. (2004). *Beyond essay length: Evaluating e-rater's performance on TOEFL essays* (TOEFL Research Rep. No. RR-73). Princeton, NJ: ETS.

Clauser, B. E., Subhiyah R., Nungester R. J., Ripkey D. R., Clyman S. G., & McKinley D. (1995). Scoring a performance-based assessment by modeling the judgments of experts. *Journal of Educational Measurement, 32*(4)*,* 397–415.

Clauser, B. E., Kane, M. T., & Swanson, D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education, 15*(4), 413-432.

Clauser, B. E., Margolis, M. J., Clyman, S. G., & Ross, L. P. (1997). Development of automated scoring algorithms for complex performance assessments: A comparison of two approaches. *Journal of Educational Measurement, 34*(2)*,* 141-161.

Cohen J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213-20.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

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

Cucchiarini, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Acoustical Society of America, 107*(2), 989-999.

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners’ fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America*, *111*(6), 2862-2873.

Cureton, E. F. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621-694). Washington, DC: American Council on Education.

Dechert, H. W. (1980). Pauses and intonation as indicators of verbal planning in second-language speech productions: Two examples from a case study. In H. W. Dechert & M. Raupach (Eds.), *Temporal variables in speech* (pp. 271-285). The Hague, The Netherlands: Mouton.

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

Deschamps, A. (1980). The syntactical distribution of pauses in English spoken as a second language by French students. In H. W. Dechert & M. Raupach (Eds.), *Temporal variables in speech* (pp. 255-262). The Hague, The Netherlands: Mouton.

ETS. (n.d.). *TOEFL iBT scores set by universities and other score users.* Retrieved September 25, 2008, from http://www.ets.org/portal/site/ets/menuitem.1488512ecfd5b8849a77b13bc3921509/?vgnextoid=031e4e63dcc85010VgnVCM10000022f95190RCRD&vgnextchannel=333bd898c84f4010VgnVCM10000022f95190RCRD.

Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., & Butzberger, J. (2000, August). *The SRI EduSpeak system: Recognition and pronunciation scoring for language learning.* Paper presented at the annual conference of Intelligent Speech Technology in Language Learning, Dundee, Scotland.

FFmpeg. (n.d.). Home page. Available from http://ffmpeg.mplayerhq.hu/

Fiscus, J., Garofolo, J., Praybocki, M., Fisher, W., & Pallett, D. (1997). *1997 English broadcast news speech (HUB-4; English)*. Philadelphia: Linguistic Data Consortium.

Frank, E., Wang, Y. S., Inglis, G. H., Holmes, G., & Witten, I. H. (1998). Using model trees for classification. *Machine Learning, 32*(1), 63-76.

Freed, B. (Ed.). (1995). *Second language acquisition in a study abroad context*. Philadelphia: John Benjamins.

Haberman, S. J. (1979). *Analysis of qualitative data: Vol. 2, new developments*. New York: Academic Press.

Hansen, L., Gardner, J., & Pollard, J. (1998). The measurement of fluency in a second language: Evidence from the acquisition and attrition of Japanese. In B. Visgatis (Ed.), *On JALT '97: Trends and transitions. Proceedings of the JALT 1997 Conference on Language*

Teaching and Learning (pp. 37-46). Tokyo: The Japan Association for Language Teaching.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing*. Upper Saddle River, NJ: Prentice-Hall.

Kaplan, R. M., & Bennett, R. E. (1994). *Using the free-response scoring tool to automatically score the formulating-hypothesis item* (ETS GRE Technical Rep. GRE 90-02b). Princeton, NJ: ETS.

Kane, M. T. (1992). An argument-based approach to validity, *Psychological Bulletin, 112*(3), 527-535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342.

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31-35.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives, 2*(3), 135-170.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed). *Educational measurement* (4th ed., pp. 17-64). Washington, DC: American Council on Education/Praeger.

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

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211-240.

Leacock, C., & Chodorow, M. (2003). C-rater: Scoring of short-answer questions. *Computers and the Humanities, 37*(4), 389-405.

Lennon, P. (1984). Retelling a story in English as a second language. In H. W. Dechert, D. Mohle, & M. Raupach (Eds.), *Second language productions* (pp. 50-68). Tübingen, Germany: Gunter Narr.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40*(3), 387-417.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*(8), 707–710.

Malvern, D. D., & Richards, B. J. (2000). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing, 19*(1), 85-104.

Manning, C., & Schuetze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Mislevy, R. J., Steinberg, L. S., Almond, R. G., & Lukas, J. F. (2006). Concepts, terminology , and basic models of evidence-centered design. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring for complex constructed response tasks in computer based testing* (pp. 15-48). Mahwah, NJ: Lawrence Erlbaum.

Moehle, D. (1984). A comparison of the second language speech of different native speakers. In H. W. Dechert, D. Mohle, & M. Raupach, (Eds.), *Second language productions* (pp. 26-49). Tübingen, Germany: Gunter Narr.

Munro, M., & Derwing, T. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition*, *23*(4), 451-468

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 47*(1), 238–243.

Page, E. B., & Petersen, N. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan, 76*(7), 561-565.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference* (2nd ed.). San Francisco: Morgan Kaufmann.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 37*(2), 257–286.

Rabiner, L., & Juang, B.-H. (1993). *Fundamentals of speech recognition*. Englewood Cliffs, NJ: Prentice Hall.

Raupach, M. (1980). Temporal variables in first and second language speech production. In H. W. Dechert & M. Raupach (Eds.), *Temporal variables in speech* (pp. 271-285). The Hague, The Netherlands: Mouton.

Read, J. & Nation, P. (2004). *An investigation of the lexical dimension of the IELTS speaking test*. Canberra, Australia: International English Language Testing System.

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations source. *Discourse Processes, 14*(4), 423-441.

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of the Intellimetric essay scoring system. *Journal of Technology, Learning and Assessment, 4*(4). Available from http://escholarship.bc.edu/jtla/

Sebrechts, M. M., Bennett, R. E., & Donald, R. A. (1991). *Machine-scorable complex constructed-response quantitative items: Agreement between expert system and human raters' scores* (ETS Research Rep. No. RR-91-11). Princeton, NJ: ETS.

Steinberg, D., & Colla, P. (1995). *CART: Tree-structured non-parametric data analysis*. San Diego, CA: Salford Systems.

Steinberg, D., & Colla, P. (1997). *CART – Classification and regression trees*. San Diego, CA: Salford Systems.

Strik, H., & Cucchiarini, C. (1999). Automatic assessment of second language learners' fluency. In *Proceedings of the 14th international congress of phonetic sciences* (pp. 759-762). Berkeley, CA: University of California, Berkeley.

Toulmin, S. E. (2003). *The uses of argument* (updated ed.). Cambridge, England: Cambridge University Press.

Towell, R. (1987). Variability and progress in the language development of advanced learners of a foreign language. In R. Ellis (Ed.), *Second language acquisition in context* (pp. 113-127). Toronto, Ontario, Canada: Prentice Hall.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83*(2), 213–217.

Williamson, D. M., Bejar, I. I., & Hone, A. S. (1999). "Mental model" comparison of automated and human scoring. *Journal of Educational Measurement, 36*(2)*,* 158–184.

Wood, D. (2001). In search of fluency: What is it and how can we teach it? *Canadian Modern Language Review, 57*(4), 573-589.

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: ETS.

Xi, X., Zechner, K., & Bejar, I. I. (2006, April). *Extracting meaningful speech features to support diagnostic feedback: A construct-driven approach to automated scoring.*

Symposium paper presented at the annual conference of National Council on Measurement in Education, San Francisco.

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education, 15*(4), 391-412.

Zechner, K., Bejar, I. I., & Hemat, R. (2007). *Towards an understanding of the role of speech recognition in non-native speech assessment* (ETS Research Rep. No. RR-07-02). Princeton, NJ: ETS.

Notes

[1] From Table 11, center column.

[2] From Table 2.

[3] Gini is the default splitting rule used in CART that typically yields the best performance. For a multiclass problem, the Gini method tends to create splits where one target class prevails. The Entropy method tends to create splits where as many score classes are evenly distributed, thus putting more emphasis on getting accurate classification of rare classes.

[4] In this evaluation we again used the ratings provided by our second set of raters. As in the scoring model experiments above, we took the second set of ratings to be more reliable.

List of Appendixes

Page

Appendix A

Rating Form for Automated Speech Features

In this table, for each feature, please rate the extent to which you agree or disagree with Statements 1–3. Please provide rationales for your rating in the cell right below each rating cell.

No	Feature name	Feature class	Dimension	1. This feature is clearly linked to a key dimension in the rubric.	2. This feature represents the feature class very well.	3. This feature represents the dimension very well.
				Strongly disagree Strongly agree →	Strongly disagree Strongly agree →	Strongly disagree Strongly agree →
1	numwds	Fluency	D	1 2 3 4 5 6	1 2 3 4 5 6	1 2 3 4 5 6

Appendix B

Sample Recognition Output (CTM File)

```
HEADER 7000100-VB123456
UTT-Power 1.0 2.0 3.0 4.0 5.0 6.0
UTT-SmoothPitch 200.0 220.0 240.0 220.0 240.0 200.0
CTM 7000100-VB123456_0 1 0.0 0.38 this 0.99
CTM 7000100-VB123456_0 1 0.38 0.85 is 0.98
CTM 7000100-VB123456_0 1 0.85 1.11 an 0.99
CTM 7000100-VB123456_0 1 1,52.2.42 example 0.94
TRAILER AM: 222111.0 LM: 63050.1
```

Appendix C

Rating Form for Evaluating the Construct Representation of Candidate-Scoring Models

1. How well do the features included in the model represent the TOEFL iBT speaking rubric?

1 2 3 4 5

Not well Moderately well Very well

2. Given the limited number of automated features available, how well does the model capture the relationships between automated features and the speaking construct?

1 2 3 4 5

Not well Moderately well Very well

3. How consistent is the model with the decision-making processes that human raters use to derive a holistic score?

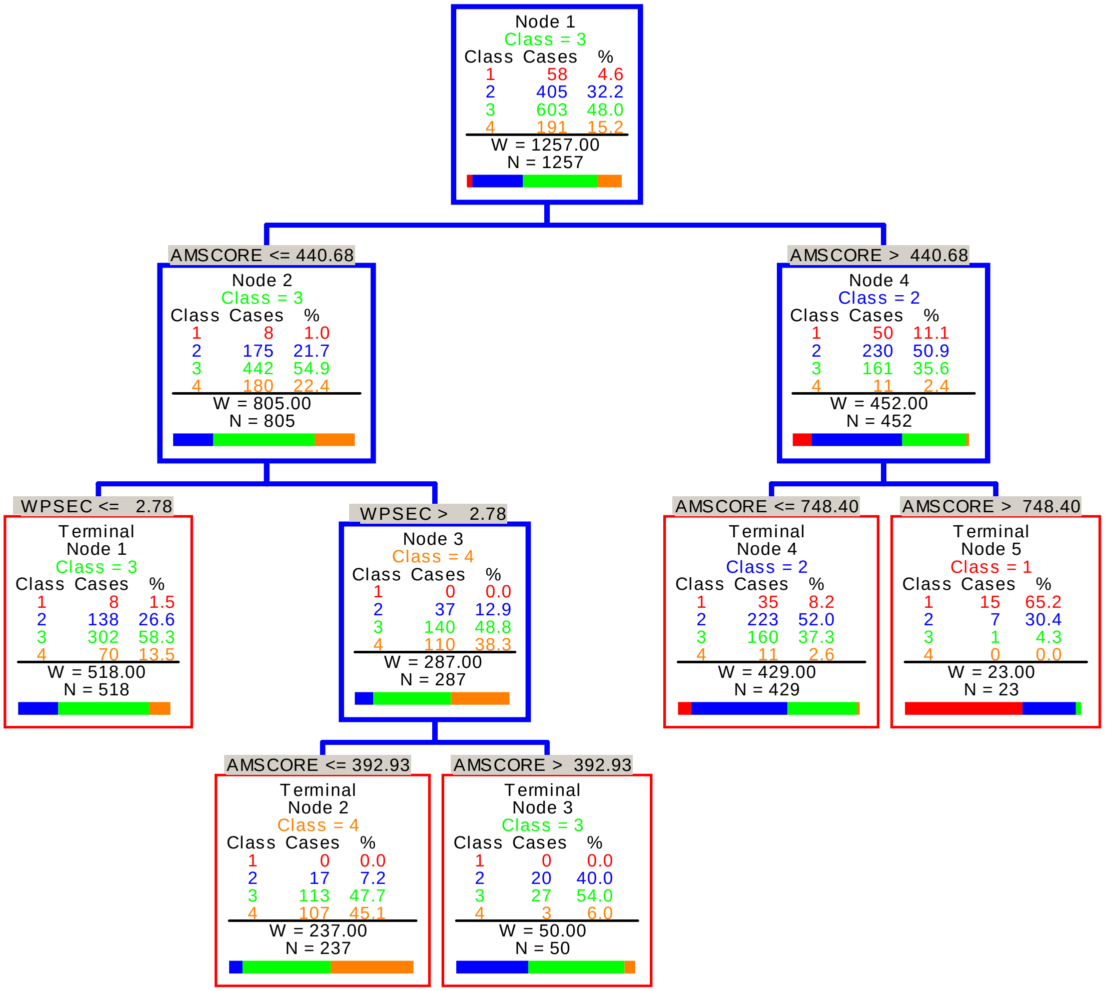
1 2 3 4 5

Not consistent Somewhat consistent Very consistent

Appendix D

The Optimal Tree for Classifying Different Score Classes

(Mixed Priors, Entropy Splitting Rule)



Appendix E
Score Report for the TOEFL iBT Speaking Practice Test

TOEFL® IBT SPEAKING PRACTICE SCORE REPORT

The purpose of this practice test is to help you prepare for the TOEFL iBT Speaking section. Performance on this test is not necessarily a predictor of how you might perform during an actual TOEFL administration because you are not taking this test under regular testing conditions. However, you are encouraged to use this practice test to get a better sense of the TOEFL iBT content and format, as well as receive scoring information on this preparation experience. ***Scores and information presented in this score report are for preparation use only and are not official test scores.***

Section	Scaled Score*	Scaled Score Range**
Speaking	22	17–27

*The Speaking section of this practice test was scored by an automated scoring system. Click here for further explanation of how this section of your test was scored.

** The Scaled Score Range represents the scores that you might expect on this TOEFL Practice Speaking test if your responses were scored by a human grader using the TOEFL iBT scoring rubric rather than SpeechRater. If your responses to this TOEFL Practice Speaking test were graded by human raters rather than SpeechRater, your score would be expected to fall within the score range provided 90% of the time.

Appendix F
How Your Practice Test Was Scored (Speaking)

Your six speaking tasks were scored by the SpeechRater program, designed as an automated scoring system for the TOEFL® Practice Speaking tests. This program uses speech recognition and processing technology to evaluate important features of your spoken responses. The SpeechRater program is currently able to analyze pronunciation, fluency and some aspects of the vocabulary and grammar of spoken responses. However, the ability of the SpeechRater program to assess the content features of a response is still limited. In general, SpeechRater scoring is based on some, but not all, of the features currently evaluated by human raters for the Speaking section of the TOEFL iBT.

To compute your Scaled Score, the SpeechRater program used all six of your responses to determine your overall spoken abilities on this test based on pronunciation, fluency, vocabulary and grammar features. SpeechRater scores for these features from the six items were combined. The total score was then converted to a score on a scale from 0–30. The reported score for the Speaking Section of the TOEFL iBT will also always be on this scale of 0–30. Because SpeechRater scoring is based on a subset of the criteria used by human raters, the Scaled Score it provides should not be considered more than *an estimate* of potential performance.

Your score report for the Speaking Section also provides you with a Scaled Score Range. The score range represents possible scores you might obtain on the Speaking section of the TOEFL iBT. Your performance on the TOEFL iBT is likely to fall within the score range provided at least 90% of the time.

Appendix G
Frequently Asked Questions About TOEFL Practice Online—Automated Scoring for Speaking

How were my responses scored?

Your responses were scored by a computer using a program specifically designed as an automated scoring system for the TOEFL Practice Online Speaking test. The SpeechRater computer program uses speech recognition and analysis technology to analyze your responses. The important features of your spoken responses that are analyzed to produce your scores are Pronunciation, Fluency, Vocabulary, and Grammar.

Why are responses to the TOEFL Practice Online Speaking test rated by a computer?

Using a computer instead of a human rater allows ETS to report your score on the Practice test within minutes.

How is my TOEFL Practice Online Speaking test score different from a TOEFL iBT Speaking score?

Both the TOEFL Practice Online Speaking test and TOEFL iBT Speaking test report your score on a scale of 0 to 30. However, the scores differ in two distinct ways: (1) the Practice test was scored by a computer rather than by a human rater and (2) the Practice test was scored by evaluating some, but not all, of the features evaluated by human raters for the TOEFL iBT Speaking test.

How is SpeechRater automated scoring different from human rater scoring?

SpeechRater scoring is an automated prediction of a score a human rater would assign for the same response. The score is produced by combining the evaluation of several important features of each response (pronunciation, fluency, vocabulary, and grammar.) Together these features cover part of the scoring criteria used by human raters to score TOEFL iBT Speaking test. Human raters scoring TOEFL iBT evaluate each response in the areas of Delivery (pronunciation, rhythm, intonation, rate of speech, pause structure, fluidity), Language Use (vocabulary and grammar), and Topic Development (content and coherence.)[1] While

[1] More information about the TOEFL iBT Scoring Rubrics is available at the ETS website: http://www.ets.org.

SpeechRater analyzes features of Delivery and Language Use, it is currently limited in its evaluation of Topic Development features. In general, SpeechRater scoring is based on a subset of the criteria currently evaluated by human raters for TOEFL iBT Speaking test.

How do I get the most benefit from my TOEFL Practice Online Speaking score?

A number of circumstances could cause this score to be a poor predictor of the scores human raters would have assigned your responses. Speakers who respond casually or carelessly to the speaking tasks, such as by responding with rehearsed speech that is not based on the speaking task, by responding in your native language, by reading from texts or notes, or by providing partial or incomplete responses are more likely to receive inaccurate scores. On the other hand, speakers who respond to the tasks seriously, as they would during a TOEFL iBT testing situation, are more likely to obtain a score reflecting their performance on practice speaking tasks that are similar to TOEFL iBT tasks.

How does computer scoring work?

Your responses to each of the six speaking questions are recorded and sent to ETS where they are analyzed by SpeechRater, the automated speech recognition and analysis software system designed especially for TOEFL Speaking Practice test scoring. In developing the system, numerous responses to TOEFL Speaking Practice test questions were processed to establish a scoring model that defined the relationship between scores by human raters and the features of your responses (Pronunciation, Fluency, Vocabulary and Grammar) that are analyzed by SpeechRater. This scoring model was then slightly modified to reflect language experts' judgments about how the score and important features of a response should be related. This modified scoring model is used to determine your score on each of the six tasks, which are then added together. The final score is converted to a 0-30 scale.

Has the computer scoring been reviewed by experts?

Language learning specialists and testing experts, both internal and external to ETS, were invited to advise and participate in selecting the speech features used to compute automated scores. The features currently used in the computer scoring were determined by these experts to represent important aspects of the scoring rubrics used by the TOEFL iBT Speaking test. The resulting combination of features was determined to provide an acceptable overall evaluation of spoken responses for use in the TOEFL Practice test.

Is my TOEFL Speaking Practice test score a prediction of my score on TOEFL iBT Speaking?
While these practice materials are designed to help you better prepare for the TOEFL iBT Speaking, the score you receive on the practice materials may not be the same as what you would receive on the TOEFL iBT Speaking. Studies have not yet been completed to compare performance on the TOEFL Practice test and the TOEFL iBT Speaking test. Until such studies are completed it will not be known how closely TOEFL Speaking Practice scores predict TOEFL iBT Speaking scores.