Measuring the Power of Learning.®

Approaches to Automated Scoring of Speaking for K–12 English Language Proficiency Assessments

ETS Research Report No. RR-17-18

Keelan Evanini • Maurice Cogan Hauck • Kenji Hakuta

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Policy Information Report and ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Approaches to Automated Scoring of Speaking for K–12 English Language Proficiency Assessments

Keelan Evanini,[1] Maurice Cogan Hauck,[1] & Kenji Hakuta[2]

1 Educational Testing Service, Princeton, NJ
2 Stanford University, Stanford, CA

This report is the fifth in a series concerning English language proficiency (ELP) assessments for English learners (ELs) in kindergarten through 12th grade in the United States. The series, produced by Educational Testing Service (ETS), is intended to provide theory- and evidence-based principles and recommendations for improving next-generation ELP assessment systems, policies, and practices and to stimulate discussion on better serving K–12 EL students. The first report articulated a high-level conceptualization of next-generation ELP assessment systems (Hauck, Wolf, & Mislevy, 2016). The second report addressed accessibility issues in the context of ELP assessments for ELs and ELs with disabilities (Guzman-Orth, Laitusis, Thurlow, & Christensen, 2016). The third report focused on critical policy and research issues of summative ELP assessments that state use for accountability purposes (Wolf, Guzman-Orth, & Hauck, 2016). The fourth report dealt with one of the major uses of ELP assessments — the initial identification and classification of ELs (Lopez, Pooler, & Linquanti, 2016). The present report discusses approaches to using automated scoring technology for the evaluation of student-spoken responses on K–12 ELP assessments. As many states have begun to use computer-based ELP assessments, there is a growing interest in automated scoring of spoken responses to increase the efficiency of scoring. This report delineates major areas to consider in using automated speech scoring for K–12 ELP assessments (i.e., assessment construct and task design, scoring and score reporting, and artificial intelligence (AI) model development and test delivery) and makes recommendations for states on how to evaluate these considerations and determine a path forward.

Keywords Automated speech scoring; K–12 English language proficiency assessment; artificial intelligence scoring; standards-based assessment; English learner speaking proficiency

doi:10.1002/ets2.12147

The purpose of this report is to inform those who are invested in and/or responsible for designing and implementing kindergarten through 12th-grade English language proficiency (ELP) assessments in schools in the United States about the potential benefits of and considerations related to using automated scoring technology (often referred to as artificial intelligence [AI] scoring) to evaluate the spoken language proficiency of K–12 students who are or may be classified as English learners (ELs).[1] Automated scoring technology is a rapidly evolving area with both significant capabilities and important limitations; it is the goal of this report to describe its capabilities and limitations in a manner that will help to support stakeholders in making decisions about whether and how to include AI scoring technology in K–12 ELP assessments.

Since the 2002 reauthorization of the Elementary and Secondary Educational Act (ESEA) as the No Child Left Behind Act (2002), states have been required to administer annual assessments of ELP and to include speaking in these assessments (along with listening, reading, and writing) to all EL students. The 2015 reauthorization of the ESEA as the Every Student Succeeds Act (ESSA, 2015) has continued this requirement for speaking to be assessed, and the current generation of K–12 ELP standards, for example, ELP or English-language development standards from the California Department of Education, the English Language Proficiency Assessment for the 21st Century consortium, and the WIDA Consortium, have been written to correspond to the challenging linguistic demands of college- and career-ready (CCR) standards such as the Common Core State Standards and the Next Generation Science Standards.

The speaking construct is an essential component of the overall construct of ELP; in the current generation of K–12 ELP standards, students are expected to be able to speak in order to fulfill challenging functions like presenting academic content information and justifying an argument. In addition to the central place of speaking in the ELP construct, recently

Corresponding author: K. Evanini, E-mail: kevanini@ets.org

developed CCR standards, such as the Common Core State Standards—English Language Arts (e.g., National Governors Association Center for Best Practices, 2010) have also increased interest in assessing the speaking and listening skills of all students. Although the focus of this report is on using automated scoring for ELP speaking assessments, the principles discussed are also applicable to other speaking assessments. We would argue that oral language assessments have the potential to be of value to all students, both because these skills are an integral part of the standards and because information obtained from speaking assessments complements the information that can currently be obtained from literacy-based assessments.

At the same time, speaking remains uniquely challenging to assess. Because of the challenges of capturing and transmitting spoken responses, traditional models of speaking assessment have typically relied on either a low-tech, in-person interview model (in which student responses are rated "in the moment" by a trained administrator) or on a technology-enabled model in which student-spoken responses are digitally captured and distributed to raters via a web-based scoring interface.

Although these approaches have been used for years to evaluate students' speaking skills, each of the two approaches has its limitations. The in-person interview model creates a considerable burden on teachers, who must undergo fairly extensive training and then must administer and score each student's assessment (consider the time required, for example, in California, where well over one million K–12 ELP speaking assessments are administered individually each year). In addition, this model provides limited avenues to check the reliability of the ratings and provide ongoing training to teachers after the initial training session. Rating via a distributed network of human raters removes the scoring burden from teachers and allows for reliability checks and ongoing rater training. However, this approach still requires a considerable amount of time to generate scores and has per-student costs that do not decrease with higher volumes of students.

In recent years, there has been increased interest in using automated scoring within K–12 ELP assessments. One reason for this is that as the technology base within schools has become more robust, the feasibility of computer-based assessments has improved. In addition, recent research has led to improvements in the accuracy of state-of-the-art speech recognizers, which leads to an improved ability of automatic speech scoring systems to effectively measure the broad range of English speaking skills that ELs need to acquire. Although this report focuses on automated scoring of speech, many of the principles discussed—including both the importance of ensuring adequate construct coverage and the need to make appropriate plans in terms of both infrastructure and timelines for implementation—also apply to the automated scoring of written responses. The following section presents the main points that stakeholders should take into account while considering the use of automated speech scoring for a K–12 ELP assessment, and the final section concludes with some practical recommendations.

Major Areas to Consider Concerning the Use of Automated Speech Scoring for K–12 ELP Assessment

In considering the use of automated speech scoring for a K–12 ELP assessment, it is crucial to find a solution that meets the following three criteria:

- It supports assessment of the speaking construct in a valid, fair, and reliable manner.
- It works practically and meets the needs of students, administrators, and score users.
- It makes efficient use of available human, technological, and financial resources.

This section addresses a range of factors that contribute to these three criteria by discussing assessment construct and task design considerations that determine whether an assessment using automated speech scoring can be valid, fair, and reliable; scoring and score reporting considerations that influence how students, administrators, and score users interpret and make use of the assessment results; and AI model development and test delivery considerations that have a practical impact on available resources.

Assessment Construct and Task Design Considerations

The most important decisions are those about the aspects of the speaking construct, that is, the test taker's knowledge, skills, and abilities for which the assessment should elicit evidence. For K–12 ELP assessments, these decisions are made based on the applicable standards for English-language development for ELs, which have been developed to correspond[2]

Table 1 California English Language Proficiency Standard Covering an English Learner's Ability to Support Opinions in Spoken English

Standard	Supporting Opinions (Kindergarten Standard Part I, C.9)
Emerging	Offer opinions and provide good reasons (e.g., *My favorite book is X because X*) referring to the text or to relevant background knowledge
Expanding	Offer opinions and provide good reasons and some textual evidence or relevant background knowledge (e.g., paraphrased examples from text or knowledge of content)
Bridging	Offer opinions and provide good reasons with detailed textual evidence or relevant background knowledge (e.g., specific examples from text or knowledge of content)

to the content area CCR standards that apply to all students. For example, Table 1 lists one of the California English Language Development Standards for ELs in kindergarten (California Department of Education, 2012).

A task type designed to elicit evidence of the student's ability to meet this standard might consist of a prompt in which students are asked to make a choice between two stated grade-appropriate options in a school-based or academic context (e.g., for first-grade students, choosing between two playground activities; for high school students, choosing between two approaches to an academic assignment) and then provide reasoning to support their choice.

To provide a valid assessment of a student's spoken response to this task type based on the standard included in Table 1, an automated speech scoring system would need to be able to (a) accurately understand the words spoken by the student using automatic speech recognition, (b) recognize the opinion and supporting reason(s) provided by the student, and (c) evaluate the relevance and appropriateness of the reason(s) provided by the student. If the automated scoring system is not able to perform these three functions, then the score it produces for a student's response to this task will not be based on evidence for the construct that the task is intended to elicit and will thus have reduced validity. Of course, a valid automated scoring system would also need to be able to evaluate other aspects of the speaking construct, such as pronunciation, fluency, and grammar, that an EL must master to provide a successful response to this task. The discussion here focuses specifically on automated evaluation of the test taker's opinion and supporting reasons, because that will likely be the most challenging task for an AI scoring system.

Therefore an automated speech scoring approach can succeed only if it is applied to speaking tasks that can be effectively assessed by current automated scoring capabilities. For example, speaking tasks can range along a continuum from heavily restricted (such as reading a sentence aloud) to completely spontaneous (such as providing a personal narrative on a topic of the student's choosing). Automated scoring (and the automated speech recognition on which it relies) tends to be most effective on more restricted tasks, in which the content of the student's response is predictable based on the content of the prompt, because these task types are typically designed to elicit aspects of the construct that can be readily assessed by state-of-the-art automated scoring systems. Table 2 provides a summary of several task types for which automated speech scoring research results are available; they are presented to reflect this continuum, ranging from the most restricted at the top of the table to the least restricted at the bottom. In addition to providing a brief description of the task type, the table also lists the main aspects of the construct that can be assessed using automated speech scoring technology as well as the readiness of the automated scoring capability for each task type based on current performance of automated scoring models for these tasks. The three levels of readiness in Table 2 are defined as follows:

- *mature*: The automated speech scoring system provides solid construct coverage and empirical performance (matching human–human agreement standards).
- *developing*: The automated speech scoring system covers some, but not all, crucial aspects of the construct and/or empirical performance is close to, but does not match, human–human agreement standards.
- *in initial stages of development*: Initial research and development has been conducted on the automated speech scoring system, but many crucial aspects of the construct are not addressed and/or empirical performance is far from human–human agreement standards.

As shown in Table 2, mature automated speech scoring systems exist for task types that elicit restricted speech. These systems provide broad coverage of the delivery aspects of the speaking construct, including pronunciation, fluency, and intonation, and typically produce scores that correlate with human scores at or above human–human agreement rates. Automated scoring systems can therefore be used for tasks of this nature with a high degree of confidence in their validity

Table 2 Sample Speaking Task Types That Have Been Researched Using Automated Speech Scoring

Task type	Description	Main aspects of the speaking construct assessed	Readiness of automated scoring	Reference
Read aloud	Test taker reads aloud a selection of printed text (word list, sentence, paragraph, etc.)	Pronunciation, fluency, intonation	Mature	Evanini et al. (2015)
Repeat aloud	Test taker hears a short recording of spoken English (usually a phrase or short sentence) and repeats it aloud	Pronunciation, fluency, intonation	Mature	Cheng, D'Antilio, Chen, and Bernstein (2014)
Keyword sentence completion	Test taker produces a complete sentence based on keywords using a predefined template	Fluency, intonation, grammar	Mature	Zechner et al. (2014)
Structured narrative	Test taker narrates a story based on a series of pictures or by retelling a story provided to the test taker in the stimulus materials	Fluency, grammar, vocabulary, discourse coherence, content	Developing	Hassanali, Yoon, and Chen (2015); Evanini et al. (2015); Somasundaran, Lee, Chodorow, and Wang (2015)
Providing an opinion	Test taker provides an opinion and supporting reasons in response to an open-ended prompt	Fluency, grammar, vocabulary, discourse coherence, content	Developing	Xie, Evanini, and Zechner (2012)
Listen–speak	Test taker listens to a presentation of academic content and provides a summary of the key points; images and/or keywords may be available for support	Fluency, grammar, vocabulary, discourse coherence, content	Developing	Xiong et al. (2013)
Conversation-based assessments	Test taker participates in an interactive conversation with virtual interlocutors	Interactive communication strategies, pragmatic competence	In initial stages of development	Evanini et al. (2014)

and reliability. On the other hand, automated scoring systems are not yet fully mature for most task types that elicit spontaneous speech, because their ability to assess the content of the spoken response is still developing. However, targeted research and development into task-specific content features has been shown to increase the validity and reliability of automated speech scoring for spontaneous speech. For example, Xiong, Evanini, Zechner, and Chen (2013) demonstrated that the addition of content features specifically designed for a content-based academic speaking task for middle school ELs improved the correlation between the automated scoring system and human scores from .62 (with no content features) to .66 (with content features), compared to a human–human agreement of .72. This result suggests that automated scoring capabilities may be able to move from developing to mature with additional research and development.

The strengths and limitations of automated speech scoring across the continuum shown in Table 2 create some interesting challenges for designing speaking tasks and assessing the speaking construct. To increase the likelihood of success for an automated speech scoring system, it is advantageous to constrain the content of the student's response in some way.

However, the speaking section of an ELP assessment, as a whole, must appropriately sample the domain of speaking as defined in the ELP standards that the assessment is designed to assess.

Current ELP standards, written to correspond to CCR standards, often emphasize the aspects of language proficiency that are required to express higher order skills involving critical thinking, as demonstrated by the Supporting Opinions standard shown in Table 1. These aspects of speaking proficiency are best assessed via assessment tasks that elicit spontaneous speech and match the target language use of the classroom. Therefore it is crucial to find a balance between these constraints when designing tasks to be included in an ELP assessment that employs automated scoring. The overall goal must be to ensure that scores on the speaking domain, on the whole, support the claims that the assessment wishes to make regarding the speaking ability of students.

Scoring and Score Reporting Considerations

Automated speech scoring technology can be used in a wide variety of ways to enhance English-language learning and assessment for K–12 ELs, and several aspects of scoring and score reporting need to be taken into account, depending on how automated scoring is applied.

The main decision that needs to be made is whether automated scoring will be used as the sole scoring mechanism for a speaking assessment (a *fully automated* approach) or in conjunction with human rating (a *hybrid* approach). A fully automated approach has great appeal, as it assumes that, once the scoring models have been built and appropriately tested, scores can be reported very rapidly (either instantly or within a small number of hours), and incremental costs associated with scoring additional test takers will be very modest. However, unless the construct can be assessed appropriately by task types for which automated scoring is fully mature and/or the decisions to be made based on speaking scores are low stakes, it is very likely that some degree of human involvement will be necessary to ensure the validity of scores.

Hybrid approaches can take a range of forms. The approach with the least human involvement is one in which humans are available as a backup to the automated system and spoken responses are routed to human raters when flagged by the automated scoring system as anomalous or difficult for the system to score (e.g., because of background noise or disruptions in the audio signal, non-English responses, and responses that are off topic). There are also hybrid approaches in which each spoken response is scored by both a human rater and the automated scoring system; these approaches can be grouped into two general categories based on whether the automated score is used as a *contributory score* (in which the final score is based on a weighted combination of human and automated scores) or a *confirmatory score* (in which responses that receive discrepant scores from the human and automated sources are sent to a second human rater for adjudication). A further variant is a complementary hybrid scoring approach in which each spoken response is scored by either a human rater or the automated scoring system and the score for the speaking section is computed based on a weighted combination of the scores from all of the responses. This configuration can be used in instances when the automated scoring system can produce valid scores for a portion of the items in the speaking section (e.g., items that elicit restricted or partially restricted speech) but lacks a sufficient number of construct-relevant features to score the others (e.g., items that elicit spontaneous speech).

In these hybrid approaches, the practical benefits of automated scoring with respect to cost savings and score turnaround time are not as striking as they are in the fully automated model, but the automated scoring technology does considerably reduce the amount of human labor needed. Furthermore, some studies have shown that the combination of human and automated scores can result in more reliable scores than using either human or automated scores alone (e.g., Breyer, Rupp, & Bridgeman, 2017, demonstrated that a combination of human and automated scores for an essay writing task in a graduate-level admissions test produced more reliable scores than either score in isolation), and a hybrid approach also helps to ensure that speaking scores are valid for their intended purposes.

In addition to deciding whether to take a fully automated approach or a hybrid approach, further decisions should be made about how the automated scoring system will be used in the process of producing the scores that are reported for the assessment. For example, the automated scoring system could report *holistic* scores only or also report *analytic* scores corresponding to different aspects of the speaking construct (such as pronunciation, fluency, and grammar). Furthermore, the speaking scores can be reported at the *item level* or at the *section level*: Item-level scores can be for providing more descriptive feedback about students' abilities, whereas a summative assessment used in making high-stakes decisions that

Table 3 Potential Use Cases for Automated Speech Scoring

Use case	Stakes associated with decisions made based on speaking score	Characteristics of scores
In-class, interactive computer-assisted language learning application	• Low stakes • Diagnostic, formative uses	• Item-level scores used for immediate feedback • Analytic scores used to provide information about student's proficiency in specific linguistic areas and to route the student to appropriate learning activities
Interim/benchmark assessments	• Medium stakes • Speaking score contributes to judgment of what instructional support a student previously identified as an EL receives	• AI scoring engine produces item-level holistic scores that are combined into a section-level score that is reported to the district
Identification assessment	• Medium to high stakes • Speaking score contributes to classification of student as EL or not EL	• Hybrid scoring approach: several items are automatically scored, and only a small number of less constrained tasks are locally scored "in the moment" by local educators (so training and scoring burden on local teachers is reduced) *or* • Fully automated approach may be possible if the score report can be designed so that speaking scores contribute to a broader oral skills (listening + speaking) score or to a simple overall score rather than a distinct speaking score
Annual summative assessment for accountability	• Medium to high stakes • Speaking score contributes to determining student progress and potential exit decision	• Hybrid scoring approach: holistic scores are produced by the AI scoring engine for each relevant item, which are then combined with the human scores on the other items to produce a section-level score that is reported to the score users

Note. AI = artificial intelligence; EL = English learner.

uses a hybrid scoring approach would typically report a single section-level score for the student's speaking proficiency based on a combination of the item-level automated and human scores.

Table 3 illustrates the preceding points by listing a few potential use cases for deploying automated speech scoring in a K–12 context, stakes typically associated with decisions made based on the speaking score in each use case, and the characteristics of the automated scoring approaches that would typically be used. Table 3 is organized from use cases that make automated scoring easier to ones in which successful application of automated scoring is more challenging; we note that the final two rows in the table (identification assessment and summative assessment) represent the two functions required by federal law[3] and so are likely to be of most compelling interest.

AI Model Development and Test Delivery Considerations

Practical considerations related to AI model development and test delivery play an important role in planning to deploy automated speech scoring capabilities in standardized assessments. This section provides an overview of some of the most important of practical considerations along with recommendations about how they should be addressed. The section concludes with a summary of the major cost drivers associated with using automated speech scoring in an operational speaking assessment.

AI Model Development: Speech Recognition and Scoring Models

Two specific components need to be developed to effectively score spoken responses: the *speech recognition models* and the *scoring models*. The requirements for these two components are described in more detail in the following pages.[4]

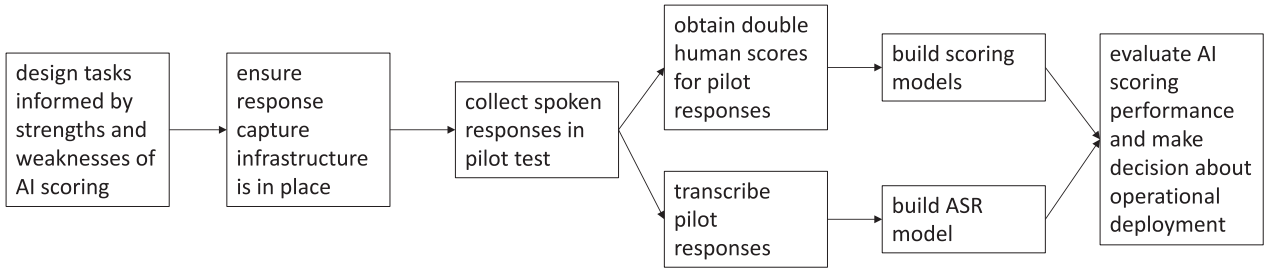

Figure 1 Schematic of the steps required for developing an automated speech scoring system. ASR = automated speech recognition; AI = artificial intelligence.

The speech recognition engine consists of an *acoustic model*, which is a statistical model of the acoustic characteristics of the speech to be processed, and a *language model*, which is a statistical model of the words and word sequences that can be recognized. To train these two components of the speech recognition engine so that it can obtain optimal performance, it is typically necessary to obtain spoken responses in a pilot deployment of the assessment with participants whose demographics match the anticipated operational population. The specific amount of data that needs to be collected in a pilot administration of an assessment to develop high-performing speech recognition models depends on factors such as the type of tasks included in the assessment and the demographic profiles (age, native language background, English speaking proficiency, etc.) of the expected test takers. However, a rough estimate of the minimum target for appropriate performance is 200 to 300 hours of speech. Obtaining accurate speech recognition results is challenging for young children's speech, and it is particularly challenging for young ELs because their speech exhibits a great amount of variability and little research has been conducted to date on automated speech recognition for this population (as a comparison, Evanini, Heilman, Wang, & Blanchard, 2015, reported speech recognition accuracy rates of approximately 70% for spontaneous speech from a middle school population of speakers with English as a foreign language, whereas the state-of-the-art accuracy rates for spontaneous speech produced by native speaker adults is approximately 94%; Xiong et al., 2016). Therefore additional training data may be required to develop a speech recognition engine with adequate performance for younger ELs in Grades K–5. In addition, developing a high-performing speech recognition engine capable of handling ELs with a wide variety of native languages (L1) is more challenging than developing an engine that is tailored to ELs with a specific L1. This means that states with large numbers of L1 backgrounds among their EL students may need to obtain a larger number of pilot responses to train a speech recognition engine that will perform well for all students. Finally, comprehensive analyses of subgroup differences in the automated scoring results should be conducted to ensure that EL students from underrepresented L1 backgrounds are being scored fairly.

The scoring model is a statistical model that generates a score for a spoken response based on a set of linguistic features that are extracted from the audio file and the speech recognition result. For a new assessment with new task types, it is necessary to obtain a sample of human-scored responses that can be used for training the statistical parameters of the scoring model. Again, the number of responses required for training the scoring model for each task type can vary depending on several factors, including the score range (e.g., 0–4) and the score distribution (balanced, skewed, etc.); a rough estimate is 1,000 responses per item. In addition, it is necessary also to collect scored responses that can be used to evaluate the empirical performance of the automated scoring system. These responses should be double-scored by independent human raters (so that automated scoring performance can be compared with human–human agreement); a rough estimate of the minimum number required is 500 responses per item.

After the targeted number of spoken responses has been collected in a pilot administration of the assessment, a sufficient amount of time needs to be included in the operational deployment schedule for building and evaluating the speech recognition and scoring models. Depending on the complexity of the data and the speech-processing infrastructure that is used, this work has typically required a minimum of 3–6 months in prior implementations. In some cases, it may be necessary to conduct additional foundational research to develop new features to expand the construct coverage of an automated scoring system prior to training scoring models. For example, to assess argumentation strategies, this requirement would extend the amount of scoring model development time required. Figure 1 summarizes the main steps in developing an automated speech scoring system that need to be scheduled into the implementation plan and accounted for in the budget.

Test Delivery: Digital Speech Capture

Automated scoring of spoken responses of course requires a mechanism for digitally capturing those responses. The following points are important in designing such a mechanism:

- *test delivery platform*: The most common delivery platforms for assessments that capture digital speech files are desktop computers and tablets; however, it is also possible to decouple the speech capture from the delivery of other sections of the assessment by capturing the spoken responses through another medium, such as the telephone.
- *audio quality*: To ensure the highest possible performance of the speech recognition engine, it is important to capture high-quality audio. This means that headset microphones are preferred over microphones built in to the computer and that background noise should be reduced as much as possible.
- *standardization*: It is important that the speech samples all be collected using a standardized procedure to ensure that the audio characteristics are similar; this means that the test delivery hardware components, such as headset microphones and computers, should be as standardized as possible across all test takers.

Because digital speech capture is a prerequisite for any computer-based speaking test, regardless of whether scores are provided by an automated system or by human raters, to ensure a smooth implementation of automated scoring if desired in the future, care should be taken to address these points related to the speech capture component of the assessment whenever a new computer-based testing system is implemented.

Costs

Although automated speech scoring has the potential to substantially reduce the cost of scoring a speaking test to ELs due to the reduction or removal of human scoring expenses from the budget, several additional expenses associated with automated speech scoring need to be considered in the planning process. The main up-front costs associated with implementing an automated speech scoring system are as follows:

- *test delivery hardware*: If adequate delivery hardware, such as headset microphones, does not exist in all testing locations, hardware needs to be purchased before any pilot data are collected and then must be maintained for operational testing.
- *pilot data collection*: A recruitment effort needs to be conducted to obtain students with the appropriate demographic characteristics to participate in the pilot data collection effort.
- *human scoring*: The responses collected during the pilot administration need to be provided with human scores by trained raters to train and evaluate the automated scoring models. Double-scoring is usually standard for the evaluation set so that human–machine agreement statistics can be compared to human–human agreement statistics.
- *automated speech scoring system development*: A sufficient amount of staff time needs to be included in the budget to build new speech recognition and scoring models based on the pilot data.

In addition to these up-front costs (i.e., those that would be incurred once prior to the initial launch of an automated speech scoring capability), there may be recurring operational expenses that should be included in the ongoing annual budget for the assessment; these include the following:

- *speech recognizer license fees*: If a commercial speech recognizer is being used, license fees will need to be paid, either on a per-submission basis or as a flat fee over a fixed time window.
- *IT infrastructure*: Sufficient computer processing capabilities (in terms of RAM and CPU) must be in place to score the responses within the specified score turnaround time; for assessments that are administered to a large number of students over a short time at fixed intervals in a calendar year, such as summative K–12 assessments, it may be necessary to have access to a large number of powerful machines (ideally via a scalable cloud-based processing infrastructure) immediately after the assessment is administered to meet the score reporting requirements.
- *human raters*: Even for a speaking assessment that uses fully automated speech scoring, it may still be desirable to have human raters score some of the operational responses on an ongoing basis as a reliability sample to monitor the performance of the automated scoring system.

Because the financial benefit of implementing an automated speech scoring system is realized when the up-front expenses are offset by the savings of not using human raters for operational scoring, automated speech scoring systems are most cost-effective for assessments that have high volumes of test takers.

Recommendations for States on How to Evaluate These Considerations and Determine a Path Forward

Given the considerable advances in the technology of automated scoring, as well as the large and growing number of K–12 EL students who need to be assessed, now is a very opportune time to consider the potential of using automated scoring to assess the speech of K–12 EL students. This report has presented several factors that need to be taken into consideration before an automated scoring approach is implemented for a given assessment. Potential users of automated speech scoring should carefully weigh all of these factors for a specific use case before making a decision.

Among these factors, considerations about the validity of the assessment should be first and foremost. As discussed in the Assessment Construct and Task Design Considerations section, automated speech scoring capabilities are able to assess some areas of the speaking construct more fully than others. This means that automated scoring needs to be taken into consideration during the test design phase so that the test can include tasks that elicit evidence for the targeted knowledge, skills, and abilities in a way that is compatible with state-of-the-art automated speech scoring capabilities. It is typically much more challenging to apply an automated speech scoring system in an assessment that has already been designed without automated scoring in mind, because the system may not be able to provide valid scores for all task types. To provide full construct coverage for a given set of ELP standards, a hybrid approach, in which responses to some tasks are scored by human raters and responses to other tasks are scored by an automated system, is likely to be optimal.

From a practical perspective, potential users of automated speech scoring should carefully consider all financial aspects associated with using automated scoring in a given assessment. The AI Model Development and Test Delivery Considerations section has outlined the main categories of up-front and recurring costs that are typically associated with using automated speech scoring in K–12 ELP assessments, and potential users should consider how each of these cost drivers will be affected in a particular use case. In addition, many factors influence the magnitude of cost reductions that can be realized through introducing automated scoring, such as the volume of students tested on an annual basis, the presence of existing capabilities for digital voice capture, the procedures used for training educators to score spoken responses, and the administration of the speaking tests. These factors should all be considered together in a cost–benefit analysis when making decisions about implementing automated speech scoring for a given assessment.

Finally, potential users of automated speech scoring should consider not only the most obvious benefits of automated scoring, such as cost savings and increased score turnaround time (in comparison to distributed, centralized human scoring), but also some of the less readily apparent benefits. For example, as mentioned in the Scoring and Score Reporting Considerations section, studies have shown that using human and automated scoring together in a hybrid approach can result in scores that are more reliable than either human or automated scores alone. Another benefit of introducing an automated speech scoring system is the possibility of providing detailed information about a student's speaking proficiency in real time to teachers and students for use as formative feedback. In addition, automated scoring systems can be applied in a consistent manner across different forms of an assessment and can increase comparability of scores. Furthermore, human scores can be adversely affected by factors such as fatigue, bias, and mental state; these subjective factors can be eliminated through using automated scoring. A final example of an ancillary benefit of introducing automated speech scoring is that having in place a procedure for digital voice capture can provide additional benefits for learning and ongoing monitoring of student progress (e.g., automated scoring can support student self-assessment, peer review, and longitudinal comparison of students' developing speaking skills).

As the state of the art in the fields of automated speech recognition and automated speech scoring continues to advance, the ability of automated speech scoring systems to provide a valid assessment of K–12 EL speaking proficiency across a wide variety of task types is expected to continue to increase. Therefore potential users of the technology should routinely reevaluate decisions about the appropriateness of automated speech scoring for a given assessment. In addition, automated systems are currently being developed to assess aspects of spoken language that are relevant to K–12 standards but that go beyond the constructs typically measured in ELP assessments, such as collaborative problem solving (Bassiou et al., 2016) and interactive speaking (Evanini et al., 2014). Educators and administrators should maintain an awareness of these developments, because these next-generation automated assessment capabilities will be ready for use in formative assessments and learning applications in the near future.

Notes

1 See also the earlier reports in this series (i.e., Guzman-Orth et al., 2016; Hauck et al., 2016; Lopez et al., 2016; Wolf et al., 2016).
2 Although ESSA calls for standards to "align," we are here following the English Language Proficiency Development Framework (Council of Chief State School Officers, 2012) in using the term "correspond."
3 Federal requirements about identifying EL students (i.e., identification) and assessing the progress of their ELP attainment (i.e., accountability) are detailed in the Title I accountability and Title III reporting requirements of ESSA; see also the "Dear Colleague Letter" (U.S. Department of Justice & U.S. Department of Education, 2015).
4 In addition, it may also be necessary to develop a method of filtering out spoken responses that may not receive a valid score from the automated scoring system due to a problem with the response. This includes responses affected by a technical difficulty (such as static or background noise) that obscures the spoken response as well as responses that contain speech that is not relevant to the construct (such as non-English responses and uncooperative test taker behavior).

References

Bassiou, N., Tsiartas, A., Smith, J., Bratt, H., Richey, C., Shriberg, E., … Alozie, N. (2016). Privacy-preserving speech analytics for automated assessment of student collaboration. *INTERSPEECH 2016—17th Annual Conference of the International Speech Communication Association Proceedings* (pp. 888–892). Menlo Park, CA: SRI International.

Breyer, F. J., Rupp, A. A., & Bridgeman, B. (2017). *Implementing a contributory scoring approach for the Graduate Record Examination Analytic Writing section: A comprehensive empirical investigation* (Research Report No. RR-17-14). Princeton, NJ: Educational Testing Service.

California Department of Education. (2012). *California English language development standards*. Sacramento, CA: Author. Retrieved from http://www.cde.ca.gov/sp/el/er/eldstandards.asp

Cheng, J., D'Antilio, Y. Z., Chen, X., & Bernstein, J. (2014). Automatic assessment of the speech of young learners. *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 12–21). Stroudsburg, PA: Association for Computational Linguistics.

Council of Chief State School Officers. (2012). *Framework for English language proficiency development standards corresponding to the Common Core State Standards and the Next Generation Science Standards*. Washington, DC: Author.

Evanini, K., Heilman, M., Wang, X., & Blanchard, D. (2015). *Automated scoring for the TOEFL Junior Comprehensive Writing and Speaking Test* (Research Report No. RR-15-09). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12052

Evanini, K., So, Y., Tao, J., Zapata, D., Luce, C., Battistini, L., & Wang, X. (2014, September). *Performance of a trialogue-based prototype system for English language assessment for young learners*. Paper presented at the Interspeech Workshop on Child Computer Interaction, Singapore.

Every Student Succeeds Act (ESSA), Pub. L. 114-95 (2015).

Guzman-Orth, D., Laitusis, C., Thurlow, M., & Christensen, L. (2016). *Conceptualizing accessibility for English language proficiency assessments* (Research Report No. RR-16-07). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12093

Hassanali, K.-N., Yoon, S.-Y., & Chen, L. (2015, September). Automatic scoring of non-native children's spoken language proficiency. In S. Steidl, A. Batliner, & O. Jokisch (Eds.), *SLaTE 2015, Workshop on Speech and Language Technology in Education* (pp. 13–18). Retrieved from http://www.isca-speech.org/archive/slate_2015/sl15_013.html

Hauck, M. C., Wolf, M. K., & Mislevy, R. (2016). *Creating a next-generation system of K–12 English learner language proficiency assessments* (Research Report No. RR-16-06). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12092

Lopez, A., Pooler, E., & Linquanti, R. (2016). *Key issues and opportunities in the initial identification and classification of English learners* (Research Report No. RR-16-09). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12092

National Governors Association Center for Best Practices. (2010). *Common core state standards*. Washington, DC: Council of Chief State School Officers.

No Child Left Behind Act (NCLB) of 2001, Pub. L. 107-110, § 115, Stat. 1425 (2002).

Somasundaran, S., Lee, C. M., Chodorow, M., & Wang, X. (2015). Automated scoring of picture-based story narration. *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 42–48). East Stroudsburg, PA: Association for Computational Linguistics.

U.S. Department of Justice & U.S. Department of Education. (2015). *Dear colleague letter: English learner students and limited English proficient parents*. Retrieved from http://www2.ed.gov/about/offices/list/ocr/letters/colleague-el-201501.pdf

Wolf, M. K., Guzman-Orth, D., & Hauck, M. C. (2016). *Next-generation summative English language proficiency assessments for English learners: Priorities for policy and research* (Research Report No. RR-16-08). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12092

Xie, S., Evanini, K., & Zechner, K. (2012). Exploring content features for automated speech scoring. *Proceedings of the 2012 meeting of the North American Association for Computational Linguistics: Human Language Technologies* (pp. 103–111). Stroudsburg, PA: Association for Computational Linguistics.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., … Zweig, G. (2016). *Achieving human parity in conversational speech recognition*. Retrieved from https://arxiv.org/abs/1610.05256

Xiong, W., Evanini, K., Zechner, K., & Chen, L. (2013, August 20–September 1). Automated content scoring of spoken responses containing multiple parts with factual information. In P. Badin, et al. (Eds.), *Proceedings of the Workshop on Speech and Language Technology in Education* (pp. 137–142). Retrieved from http://www.isca-speech.org/archive/slate_2013/sl13_137.html

Zechner, K., Evanini, K., Yoon, S.-Y., Davis, L., Wang, X., Chen, L., … Leong, C. W. (2014). Automated scoring of speaking items in an assessment for teachers of English as a foreign language. *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 134–142). Retrieved from www.aclweb.org/anthology/W14-1816

Suggested citation:

Evanini, K., Hauck, M. C., & Hakuta, K. (2017). *Approaches to automated scoring of speaking for K–12 English language proficiency assessments* (Research Report No. RR-17-18). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12147

Action Editor: Anastassia Loukina

Reviewers: Lawrence Davis and Emilie Pooler