What and How Much Evidence Do We Need?  
Critical Considerations in Validating an Automated Scoring System

Xiaoming Xi  
Educational Testing Service

Building on Clauser, Kane and Swanson (2002), this paper illustrates how an argument-based approach can be applied to the validation of the TOEFL® iBT Speaking test which uses an automated scoring system called SpeechRater v.1.0. The paper outlines assumptions pertaining to the links between each stage in the score interpretation and decision making process. Finally, evidence needed to reject potential rebuttals against the inferences is described. By outlining the inferences underlying score interpretation, the paper shows the connections among various aspects of validity evidence and offers insights into practical issues arising in a validation process such as prioritization of different types of evidence.

INTRODUCTION

The emergence of automated scoring systems in the past two decades (see a review in Yang, Buckendahl, Juszkiewicz, & Bhola, 2002) has been accompanied by theoretical work that defines the nature and scope of validation and empirical research to validate these systems. Previous validation work has followed a piecemeal approach and addressed one or more of these three areas: (1) demonstrating the correspondence (in both agreement and reliability) between scores produced by automated scoring systems and by human scorers, (2) examining the relationship between automated scores and scores on external measures, and (3) understanding the scoring processes that automated scoring systems employ (Yang et al., 2002). These different areas of investigation could potentially contribute to an argument for using automated scoring in an assessment; however, a mechanism is needed to tie them together in a coherent manner. This mechanism should allow practitioners to determine the critical evidence needed in view of the targeted use of the automated scores, and to integrate and evaluate existing evidence to support an argument for using automated scoring in a particular learning or assessment context.

Fortunately, we have seen a few attempts in the last ten years to integrate automated scoring into the overall assessment process, or the overall validity argument for an assessment. The body of work described by Bennett and Bejar (1998) provides useful guidance for developing a valid computerized assessment for which automated scoring
has been planned from the outset. It also unveils the complexity of validation work related to automated scoring. Another body of work, initiated by Clauser, Kane and Swanson (2002), is most useful in guiding the development and synthesis of evidence to support the proposed interpretation and use of scores produced by an automated scoring system. Their approach integrates the various areas of validation reviewed in Yang et al. (2002) into a coherent argument and extends these areas to include decisions based on automated scores and consequences incurred from using automated scoring. It thus provides a working framework for weaving automated scoring into the validity argument for the whole assessment.

Building on Clauser et al. (2002), this paper illustrates the application of an argument-based approach to the validation of SpeechRater v1.0, an automated scoring system deployed for the TOEFL® Internet-based Test (TOEFL iBT) Speaking Practice test. By contextualizing the approach in a real-world application, it offers practical insights into how to prioritize the different types of evidence gathered to support validation research in light of the intended use of SpeechRater in an on-line practice environment.

CONCEPTUAL VALIDATION FRAMEWORKS FOR AUTOMATED SCORING

Bennett and Bejar (1998) noted that previous research has largely examined automated scoring in isolation from the other components of an assessment and contended that it should be seen as an integral part of the whole assessment process. While automated scoring is constrained by other aspects of the assessment process, automated scoring itself has influence on decisions with respect to other aspects of the assessment, such as construct definition, test and task design, test taker interface, and reporting methods. Bennett and Bejar proposed that the development of an automated scoring system should involve two key steps: 1) extracting and implementing relevant features, each of which evaluates an aspect of the performance; 2) combining them into a score that indicates the overall quality of performance. Further, these two steps could be manipulated to maximize construct representation and to improve the relationships between automated scores and human scores on the same test or on external criterion measures. Bennett and Bejar’s conceptual approach is most useful in driving the development of a valid computerized assessment that involves automated scoring. By seeing automated scoring as a dynamic component in a computerized assessment system consisting of interrelated components, this framework emphasizes the importance of evaluating the scoring mechanism in the context of a validity argument for the assessment. It has thus broadened the scope of validity investigations regarding automated scoring. Although the relationship of automated scoring to the overall validity argument of the assessment is not emphasized, their paper provides a foundation for the subsequent work that shifts the focus to the complete validity argument.

Based on a critical analysis of empirical validation efforts on automated scoring systems, Yang et al. (2002) proposed a validation framework that they claim to be essentially an
elaboration of the one developed by Bennett and Bejar (1998). Using this broadened validation framework as the reference point, they noted two gaps in the existing literature. The first one was the dearth of literature that conceptualized potential threats posed by the use of automated scoring for construct relevance and representation. They also highlighted the point that the consequences of using automated scoring systems should be examined as part of a validity argument, which would include an investigation of the extent to which it affects the user’s perceptions of the assessment and the way they interpret and use the scores. Although not explicitly discussed in their paper, the impact of automated scoring on teaching and learning, depending on the goals of a particular assessment, seems to be a natural expansion of the scope of consequences which is part of a validity argument (e.g. Kane, 2006).

Clauser et al. (2002) provided the most comprehensive and in-depth analysis of validity issues involved in automated scoring systems for performance-based tests, following a general argument-based approach to validating a whole assessment (Kane, 1992; 2001; 2002; 2006; Kane, Crooks & Cohen, 1999). With this approach, validation involves two stages: developing an interpretative argument and evaluating a validity argument. In the first stage, for each intended use of test scores, an interpretive argument is articulated through a logical analysis of the chain of inferences linking performance on a test to a score-based decision, and the assumptions upon which these inferences rest. The second stage involves an evaluation of the plausibility of the interpretive argument within a validity argument using theoretical rationales and empirical evidence.

This approach has not expanded the scope of validity investigations beyond that of Messick (1989), but its major strength lies in providing a transparent working framework to guide practitioners in three areas: prioritizing different lines of evidence, synthesizing them to evaluate the strength of a validity argument, and 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 to discount or reduce the impact of such threats. In applying this framework to automated scoring, Clauser et al. (2002) discussed how decisions made in developing an automated scoring system may strengthen the overall validity argument or potentially weaken it, given the particular approach used to develop the system. Their discussion focused on the potential threats to the strength of each inference in the chain that may be introduced by automated scoring, pointing to the critical areas of research that are needed to discount or reduce the threats. Although Clauser et al. (2002) may not cover all the potential validity issues introduced by automated scoring, they provide a working model for integrating automated scoring into this network of inferences leading to the intended interpretation and use of test scores.

Their working model is used as a basis in this paper for examining issues that might impact the validity of the TOEFL iBT Speaking Practice test which uses SpeechRater v1.0. In addition, it identifies the most critical inferences to be supported given the purpose of the assessment and summarizes evidence that is needed to reduce the impact of the potential threats to each inference.
THE TOEFL iBT PRACTICE ON-LINE ASSESSMENT

SpeechRater v1.0 is intended to provide instant score feedback on the TOEFL iBT Speaking Practice test. This section provides a brief overview of the purpose of the TOEFL iBT Practice assessment and the tasks and scoring rubrics of the TOEFL iBT Speaking Practice test.

The TOEFL iBT Speaking Section is designed to measure the academic English abilities of non-native speakers who plan to study at English-medium institutions for higher education. The TOEFL Practice On-line (TPO) has been made available to help prospective TOEFL iBT examinees become familiar with and better prepared for the TOEFL iBT test. Using retired operational TOEFL iBT test forms, TPO is designed to mirror the content and design characteristics of the TOEFL iBT test to the extent possible. However, unlike the TOEFL iBT test, 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 on-line delivery system and timing restrictions of TOEFL iBT. In the untimed mode, users can progress at their own pace, starting or stopping the test whenever they like and revisiting items they have completed if desired. Another important distinction between the TPO and the TOEFL iBT test is that the former allows users to receive immediate feedback on their performance to help them assess their own comfort with the TOEFL iBT test administration. In early 2006 the users of TPO were able to instantly receive scores on reading and listening sections, both comprised of multiple-choice items that are computer scored, as well as the writing section, with automated writing scores provided by e-rater® (Attali & Burstein, 2005). The scores on speaking sections were produced by human raters within five business days. As a result of substantial interest in more immediate feedback from the speaking section of the TPO, a research agenda was launched to develop and deploy an automated system for scoring the speaking sections. 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 raters. The long-term goal was to provide instructional and diagnostic feedback based on automated features in addition to providing valid and reliable total test scores. The result of this effort was the release of SpeechRater v1.0 for use in the TPO in November 2006.

The TOEFL iBT Speaking Practice test, like the TOEFL iBT Speaking Section, contains six tasks. The first two are independent tasks that ask candidates to speak about 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 English or comprehend spoken English. The remaining four are integrated tasks that engage reading, listening and speaking skills in combination to measure the communication skills typically required in campus-based situations and in academic courses. The entire test takes approximately 20 minutes. For each of the six tasks, the examinees are allowed a short time to prepare their response and then 45 to 60 seconds (the time limit varies by task type) to provide their response in a spontaneous manner.
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 Section. The raters issue a holistic score for each response on a score scale from 1 to 4 that is based on three key categories of performance: Delivery, Language Use, and Topic Development (see Xi & Mollaun, 2006 for the scoring rubrics).

A BRIEF OVERVIEW OF SPEECHRATER v1.0

SpeechRater v1.0 provides instant score feedback for the TOEFL iBT Speaking Practice test. It consists of three major components: the speech recognizer and feature generation programs, the scoring model, and the user interface. The speech recognizer and the feature generation programs are closely interrelated and can be considered as one integrated 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 speaking performance, based on various output that the speech recognizer produces, which may include words uttered, pauses, pitch, energy, etc. The second component is the scoring model that scores responses to individual tasks based on the scoring features and summarizes the scores across multiple tasks. The last component is the user interface that provides the users with the score report and advisory information about how to interpret and use the scores. Details about different components of this system are not included in this paper, but interested readers could refer to Xi et al. (forthcoming) for more information.

AN ARGUMENT-BASED APPROACH TO VALIDATING SPEECHRATER v1.0

This section illustrates the application of the argument-based approach to validating SpeechRater v1.0. As Clauser et al. noted, the use of automated scoring will not only impact 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” in Clauser et al. (2002). To position automated scoring in an interpretive validity argument for using the test scores for a particular purpose, a general description of the chain of inferences resulting in a decision based on language test scores is provided below. (For an elaboration on building a validity argument for a language test, see Chapelle, Enright, & Jamieson, 2008).

Figure 1 illustrates the mechanism under which various types of inferences can be organized conceptually to link a sample of test performance to score-based interpretations and uses. The process of establishing an inferential link involves building an informal argument. In particular, each inferential link rests on certain assumptions that need to be backed by evidence.
Each inference, if sustained, becomes the grounds for the subsequent inference in the argument. The first link from a sample of language test performance to test scores hinges on the assumption that performance on a language test is obtained and scored appropriately to yield accurate scores for the intended use (Evaluation). The second link is from an observed score to a universe or true score. The pertinent assumption is that performance on language test tasks is generalizable over similar language tasks in the universe, raters, test forms and occasions (Generalization). 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 in an alternate form or the same test on a different occasion, he/she would receive similar scores.

The third link between a universe score and an interpretation is crucial in the overall validity argument, because it bears on whether test takers’ performance on the test provides adequate evidence about their 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 (Extrapolation) and that speaking abilities and processes revealed by language test tasks vary in ways that are consistent with models of communicative competence in academic contexts (Explanation). At this link, 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. In some instances, in the absence of a strong construct theory, the extrapolation of test performance to the intended domain may sustain the link from the universe score to the score interpretation. The fourth link, utilization, connects score-based interpretations and decisions. The assumptions are that the test scores and other related information provided to users are relevant, useful, and sufficient for making intended decisions and that they promote positive effects on teaching and learning (Utilization) (Bachman, 2005).

Figure 1. Links in an interpretative validity argument
(Modified after Kane, Crooks & Cohen, 1999 and Bachman 2005)
In summary, in the process of building a clear and coherent chain of reasoning, more and more meaning is attached to a sample of test performance and the corresponding score to justify the final score interpretation and use. These different meanings are tied to having accurate scores, generalizable scores, meaningful scores, scores that indicate domain performance, scores that are useful for decision-making, and scores that have beneficial consequences.

When automated scoring is integrated into an assessment, its most immediate effects seem to be on the accuracy of the resulting scores, thus pertaining to the Evaluation inference. This is also the aspect of automated scoring that has been most heavily researched. However, the effects of automated scoring may extend beyond this and be evident through all of the subsequent inferences. At each stage described above, automated scoring may introduce enhancements to validity that human scoring may not be able to offer or pose threats to validity in ways that are not typical of human scoring.

The use of automated scoring can potentially enhance the validity argument that supports the intended use of test scores. Specifically, it allows the designer of an automated scoring system to maximize construct representation by selecting construct-relevant response features and combining them to produce scores in a way that best represents the construct (Bennett & Bejar, 1998), thus contributing to the strength of the Explanation inference. This degree of control is not possible with human scoring. In addition, an automated scoring system applies the defined rating criteria consistently. It can thus improve score generalizability and strengthen the Generalization inference by eliminating differences in rater leniency or harshness, in raters’ judgments over tasks, occasions or combinations of them.

Nevertheless, what may come with the systematic control of the construct is systematic error due to construct under-representation or construct-irrelevance. This is particularly true for 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 presents immense challenges. The tendency to extract easily quantifiable aspects of the performance due to the limitations of current speech technologies would potentially result in construct-irrelevant features or features that do not represent the full construct. In addition, given the complexity of human raters’ decision-making processes involved in rating speaking, it obviously is not an easy task to design a scoring system that adequately reflects those processes. Even a scoring solution informed by expert judgments may not be adequate in representing the intended construct, depending on the qualifications of the experts and the rigor with which the work is conducted.

The systematic error introduced by automated scoring may impact more than one of the inferences that lead to the interpretation and the use of the scores. For example, automated scoring may reduce task specificity by disproportionately capturing aspects of speech that are relatively stable across tasks, thus improving the score generalizability and strengthening the Generalization inference. However, it may have reduced the task
specificity in undesirable ways. It may compromise the explanatory power of the scores in representing the constructs by failing to include some aspects of speech that are construct-relevant but are less stable across tasks, thus weakening the Explanation inference.

Replacing human scoring with automated scoring may also change users’ perceptions of the assessment and the way they interact with the assessment tasks. Knowing that scores are produced by an automated system rather than human raters, users may also interpret and use the scores differently than they would use scores from human raters. Further, users typically perceive automated scoring as inferior to human scoring, although the latter is also prone to error. Their perceptions are sometimes misguided by some general misperceptions about automated scoring and may not be motivated by the specifics of a particular scoring system. Therefore, it is important to investigate how automated scoring may impact the Utilization inference through investigations of the aspects discussed above.

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 counter-evidence against the rebuttals. The validation efforts should focus on providing counter-evidence that discounts these rebuttals.

To build and evaluate a validity argument for SpeechRater v1.0, four basic steps are involved:

1) Clearly state the intended interpretation and use of the automated scores on TOEFL iBT Speaking Practice test;
2) Articulate the network of inferences that lead to the intended interpretation and use and the associated assumptions that will lend support to each inference if backed by evidence;
3) Identify critical rebuttals that may weaken each inference as a result of using automated scoring; and
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. This paper will address the first three steps and demonstrate the process of developing an interpretative argument.

The goal of developing SpeechRater v1.0 was to support the intended use of the product, i.e., help students better prepare for the TOEFL iBT Speaking and gauge their own readiness to take the official test. The claim we intend to support is:

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 could also choose to use more time to plan a response before starting to record it, but this option is not available for the official test. However, a candidate may be able to self-evaluate his/her readiness for the official test, knowing the conditions under which he/she has taken the practice test. A candidate could potentially choose to take the practice test under the timed mode and make his/her best effort to respond to each task as if he/she were taking the official test. Only under these circumstances would a candidate be able to assess his/her own readiness to take 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.

Further, SpeechRater v1.0 does not provide diagnostic feedback, although this is a long-term goal. It provides only a single score without any detail about why the score was obtained.

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. The crucial rebuttals that may potentially undermine the validity of these claims are also stated, associated with the inference to which each pertains. Failures to provide evidence to reject any of these rebuttals related to the critical inferences would potentially weaken the entire argument.

Guided by this framework, different lines of evidence can be organized into these five areas and synthesized to evaluate the soundness of the validity argument. Summarized below is key evidence relevant to each area that can potentially be gathered.

**Evaluation.** The relevant evidence includes the association between human and SpeechRater scores indicated by various well-established measures such as correlation and kappa. Different scoring methodologies that are used to produce SpeechRater scores
Table 1  Areas of emphasis for validity of SpeechRater v1.0 and associated rebuttals

<table>
<thead>
<tr>
<th>Inference</th>
<th>Assumptions</th>
<th>Rebuttals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Evaluation</strong></td>
<td>Automated scoring results in scores that accurately represent the quality of the performance on the practice test.</td>
<td>1. The scoring algorithm under- or misrepresents the construct or introduces construct-irrelevance so that the resulting scores are not accurate.</td>
</tr>
</tbody>
</table>
| **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 speaking performance in a manner that is consistent with theoretical predictions about 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 under-representation.  
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 construct is under-represented. |
| **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.  
3. The automated scoring system does not promote positive washback effects on English language teaching and learning.  
4. Other potential negative consequences of SpeechRater v1.0 are not anticipated or minimized. |
based on automatically extracted speech features can be compared based on the strength of the association between the model predicted scores and the human scores. The soundness of the statistical principles underlying each methodology is also an important consideration in employing a particular scoring methodology.

**Generalization.** This inference draws support from two types of evidence. One concerns the procedures for developing and evaluating the scoring models such as the adequacy of the sample size, representativeness of the sample, and absence of overlap in speakers between the scoring model training and evaluation data. The other type of evidence includes the generalizability of the SpeechRater scores across different tasks that can be estimated using Generalizability studies (Cronbach, Nageswari, & Gleser, 1963) or other established methodologies. The score generalizability estimates could be compared to those obtained for human scores and typical figures acceptable for a practice context.

**Extrapolation.** The potential evidence supporting this inference is the association between SpeechRater scores and scores on criterion measures of students’ academic speaking ability, such as faculty or English instructors’ ratings of their students’ speaking proficiency.

**Explanation.** Evidence supporting this inference is conceptual and judgmental in essence. In particular, two essential qualities of the SpeechRater scoring model need to be verified to argue that it captures aspects of performance in a manner that is consistent with theoretical predictions about speaking abilities used in an academic setting: the construct relevance and coverage of the features and the defensibility of the way they are combined. The evidence involves largely judgments of these qualities by experts who have an intimate understanding of the construct the assessment is designed to measure, the conceptual meaning of each scoring feature used, and the way the scoring features are combined through a statistical model to produce a score that indicates the overall quality of performance.

**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 score-based decision. Arguments about potential consequences of the SpeechRater v1.0 can be made based on the score report, and the advisory information communicated to the user about the limitations of the system and the intended use of the scores can be included as part of the user interface. Additional evidence may include investigations of user perceptions, e.g., to what extent the awareness of the scores being produced by a machine impacts the way a user interprets and uses the scores, as well as the impact of using automated scoring on teaching and learning practices.

The aspects of the validity argument that require full support are dictated by the intended use of the assessment scores. For example, the areas of emphasis for validating an automated scoring system intended for a practice environment may differ from those for a system employed in an assessment for high-stakes decisions. If automated scores are
What and How Much Evidence Do We Need?

intended to support high-stakes decisions, all the five inferences discussed above need to be fully supported. The Explanation inference is especially important—if the automated scoring model under- or misrepresents the construct of interest, test takers may be misguided to focus on the wrong things or omit important things in their test preparation. It may also make the assessment more vulnerable to new types of cheating and test-taking strategies that would negatively impact the trustworthiness of the scores. An automated scoring system that under- or misrepresents the construct may also incur negative washback effects on teaching and learning and hurt the credibility of the test program.

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 rather than diagnostic feedback on learners’ strengths and weaknesses in speaking, three of the five inferences particularly need adequate backing by relevant empirical or judgmental evidence: Evaluation, Generalization and 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.

Based on the validation framework discussed above, the relevant evidence pertaining to each inference can be integrated and evaluated. Then the overall strength of the validity argument can be evaluated in light of the critical inferences that need adequate backing to support the intended claims of this version of SpeechRater.

CONCLUSION

This argument-based approach to validating an automated scoring system drives researchers to consider in a systematic way what and how much evidence is needed to justify the use of an automated scoring system in an assessment for a particular purpose. This principled approach can guide us to think through the process of articulating an interpretative argument for using an automated scoring system as well as collecting and evaluating evidence to support a validity argument.

REFERENCES


