2-26-2018

Innovative Implementation of a Web-Based Rating System for Individualizing Online English Speaking Instruction

Hyejin Yang
Sookmyung Women's University

Elena Cotos
Iowa State University, ecotos@iastate.edu

Innovative Implementation of a Web-Based Rating System for Individualizing Online English Speaking Instruction

Abstract
The primary goal of computer-assisted language learning (CALL) in general, and of online language instruction in particular, is to create and evaluate language learning opportunities. To be effective, online language courses need to be guided by an integrated set of theoretical perspectives to second language acquisition (SLA), as well as by specific curricular goals, learning objectives and outcomes, appropriate tasks and necessary materials, and learners' characteristics and abilities – to name a few factors that are essential in both online and face-to-face teaching (Xu & Morris, 2007). Doughty and Long (2003) articulate pedagogical principles for computer-enhanced language teaching, which highlight the importance of exercising task-based activities, elaborating the linguistic input, enhancing the learning processes with negative feedback, and individualizing learning. Chapelle (2009) further puts forth a framework of evaluation principles that define the characteristics of tasks and materials drawing on SLA theories. Notably, she remarks that "[t]he groundwork for such evaluation projects is an iterative process of stating ideals for the materials based on the theoretical framework and providing a judgmental analysis of the degree to which the desired features actually appear in the materials" (Chapelle, 2009: 749). In other words, she calls for a judgmental analysis as pre-evaluation. With regards to online language instruction, pre-evaluation is rather challenging when it comes to individualizing learning in view of learners' characteristics and abilities, which are different in every iteration of the course

Disciplines
Computational Linguistics | Creative Writing | Educational Technology | Online and Distance Education | Theatre and Performance Studies

Comments
This chapter is published as Yang, H., Cotos, E., (2018). Innovative implementation of a web-based rating system for individualizing online English speaking instruction. In S. Link & J. Li (Eds.), Assessment across online language education, CALICO Monograph Series (Vol. 16, pp. 167–183). *CALICO*: San Marcos, TX. Posted with permission.

8 Innovative Implementation of a Web-Based Rating System for Individualizing Online English Speaking Instruction

Hyejin Yang[*†] and Elena Cotos[**]

Introduction

The primary goal of computer-assisted language learning (CALL) in general, and of online language instruction in particular, is to create and evaluate language learning opportunities. To be effective, online language courses need to be guided by an integrated set of theoretical perspectives to second language acquisition (SLA), as well as by specific curricular goals, learning objectives and outcomes, appropriate tasks and necessary materials, and learners' characteristics and abilities – to name a few factors that are essential in both online and face-to-face teaching (Xu & Morris, 2007). Doughty and Long (2003) articulate pedagogical principles for computer-enhanced language teaching, which highlight the importance of exercising task-based activities, elaborating the linguistic input, enhancing the learning processes with negative feedback, and individualizing learning. Chapelle (2009) further puts forth a framework of evaluation principles that define the characteristics of tasks and materials drawing on SLA theories. Notably, she remarks that "[t]he groundwork for such evaluation projects is an iterative process of stating ideals for the materials based on the theoretical framework and providing a judgmental analysis of the degree to which the desired features actually appear in the materials" (Chapelle, 2009: 749). In other words, she calls for a judgmental analysis as pre-evaluation. With regards to online language instruction, pre-evaluation is rather challenging when it comes to individualizing learning in view of learners' characteristics and abilities, which are different in every iteration of the course.

Individualization is part of learner fit (Hubbard, 1988), a critical concept from cognitive and psycholinguistic SLA perspectives (Chapelle, 2009) as well as from the perspective of CALL evaluation (Chapelle, 2001; Hubbard, 2006). In essence, learner fit refers to the language level and the opportunities for engagement with language under appropriate conditions accounting for learner characteristics (Jamieson, Chapelle, & Preiss, 2005). When considering learner fit, Hubbard (2006) recommends determining if the skills and the level of language difficulty (i.e., the level of grammatical, lexical, phonetic challenge) are compatible with learner variables (e.g., native language, proficiency level, learner needs) and the course objectives in the syllabus, accentuating that learner variables are individual by nature and not evident to the teacher. In that case, how can language teachers and online course designers be informed about learner variables thorough a pre-evaluation judgmental analysis? How can they make informed decisions

* Sookmyung Women's University and Yonsei University, Korea; hjyang1112@gmail.com
** Iowa State University, USA; ecotos@iastate.edu
† Both authors contributed equally to this chapter.

regarding whether or not the linguistic forms targeted in their courses are at an appropriate level of difficulty for individual learners? These questions pose a considerable practical challenge for online speaking courses.

Concerned with this challenge, we turn to curriculum-related assessment (Carr, 2011) in an attempt to leverage its underexplored capacity to strengthen the learner fit of online language teaching. Our work focuses on the need to inform language instruction in the context of oral language proficiency courses for international teaching assistants (ITAs) at a large university in the Midwest United States, which are to undergo a transition from the face-to-face to the online mode. The purpose of our mixed-methods study is to examine the potential of R-PLAT (Rater Platform), a computer-based tool for speaking assessment, to generate diagnostic evidence of the language ability of prospective individual students. We follow the theoretically-grounded argument-based validation approach (Kane, 2016), highlighted in Chapter 6 and further discussed in Chapter 12, by empirically investigating a judgmental assumption about the intended use of assessment results from R-PLAT. The results obtained from qualitative and quantitative data support the intended use of R-PLAT as a diagnostic informant for the design of online course materials and tasks that would tailor the level of language difficulty to individual needs and speaking ability. On a broader scale, this work sets the scene for assessment-enhanced development of sound pedagogical principles necessary for curriculum design of online language courses.

Curriculum-Related Assessment

Carr (2011: 6) distinguishes assessments that "are closely related to teaching or learning curriculum, and those that are not," defining the former as curriculum-related because teachers and administrators draw on specific curricula when planning and developing them. Such assessments include placement, diagnostic, progress, and achievement tests. Of these, placement and diagnostic assessments are especially suitable for individualizing curriculum planning at the pre-evaluation stage. Among the types of assessment that Carr places outside the curriculum-related domain are proficiency tests, for they are used to determine learners' level of language ability "without respect to a particular curriculum" (Carr, 2011: 8). In practice, though, proficiency tests are often used for placement into certain levels of language courses. Sometimes diagnostic information can be derived from placement tests (Fox, 2009). Therefore, despite their distinct purposes, placement, diagnostic, and proficiency tests can complementarily serve to obtain evidence to inform curriculum design with descriptive details about the language ability of individual students.

Although leveraging the potential of these assessment types is appealing and equally justifiable, a problem surfaces with regard to individual performance descriptors. It is a common assumption that scaled proficiency descriptors of different performance tests (e.g., TOEFL, IELTS, CEFR, ACTFL) have evident diagnostic potential (Jang, 2012). In other words, performance-level descriptors provide a depiction of language ability within given performance levels that can help teachers form diagnostic judgments about learners' mastery of language based on an external standard of performance. However, those descriptors are "absolute" in nature (Carr, 2011: 10). For example, ACTFL characterizes a novice's speaking ability to use language functionally as "[c]an ask highly predictable and formulaic questions and respond to such questions by listing, naming, and identifying" and "[m]ay show emerging evidence of the ability to engage in simple conversation" (ACTFL, 2012: 14). Such descriptors are not

sufficiently informative to make specific diagnostic inferences about individual learners' strengths and weaknesses in the skills assessed. That is not to say that obtaining such information is not possible, as technological advancements in computer-assisted language testing (CALT) have increasingly enabled the integration of assessment in teaching and learning (Chapelle, Chung, & Xu, 2008). In this chapter, we provide an example of how technology can interlace the connection between assessment and teaching with diagnostic information, and how systematic evidence can be gathered and evaluated under the validity argument framework.

Argument-Based Validation and Speaking Assessments for Online Teaching

In language assessment, validation is the most essential process for justifying the use and interpretations of test outcomes. The argument-based approach to validation (Kane, 2016) consists of an interpretive argument and a validity argument. The interpretive argument specifies "the proposed interpretations and use of test results by laying out the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances" (Kane, 2006: 23). The inferences include: domain description, evaluation, generalization, explanation, extrapolation, and utilization (Chapelle, Enright, & Jamieson, 2008). Each inference is authorized by an explicit warrant; each warrant, in turn, has underlying pre-evaluation assumptions that need to be investigated empirically. The validity argument, in essence, is a process for evaluating the "coherence, completeness, and plausibility" of the proposed assumptions in the interpretative argument (Kane, 2016: 202). The final chapter in this book provides an extended description of the approach.

CALT studies employing the argument-based approach have proliferated in the past few years. With regards to ITA contexts, a number of studies provide evidence supporting the validity of the interpretation and use of the TOEFL iBT® Speaking scores for ITA certification (Farnsworth, 2013; Lim et al., 2012; Wylie & Tannenbam, 2006; Xi, 2007). Another test with a speaking component, the Pearson Test of English Academic, was similarly evaluated in university contexts on the basis of an assessment use argument (Bachman & Palmer, 2010; Wang et al., 2012). The rigor and depth of these works are exemplary; however, no such studies have been conducted for the purpose of examining diagnostic evidence with an emphasis on learner fit for online language teaching.

Online language courses focused on speaking have successfully integrated technologies for various assessment purposes. Most commonly, teachers adapt to the affordances of commercial and open-source applications to enable e-assessment of students' progress and achievement. For example, Volle (2005) used voiced audio emails and MSN Messenger in an online Spanish class to measure improvement in learners' pronunciation, stress, and intonation, as well as accuracy and overall oral proficiency. Blake et al. (2008) reported on the use of Versant, a phone-delivered automated speaking test, to assess students' oral language proficiency in the final weeks of distance learning, hybrid, and face-to-face Spanish courses. Levy & Kennedy (2004) utilized audio-video conferencing tools for enhancing learners' of Italian focus on language form and for ongoing formative assessment, which is conceptually close to diagnostic assessment in that both aim to inform differentiated instruction (Nichols, Meyers, & Burling, 2009). However, to our knowledge, no computer-based assessment of speaking has been used as a pre-evaluation diagnostic measure that would identify error patterns and discrepancies from expected performance to tailor online teaching to individual learner needs. Moreover, interpretive

arguments for curriculum-related assessments in the context of ITA language instruction have not been articulated.

The Study

Assumption and Research Questions

This study centers on the *evaluation* inference in the interpretive argument for using R-PLAT in order to inform the transition from face-to-face to ability-tailored online ITA speaking courses. Because test scores are to be interpreted in relation to this domain, we conform to the definition of evaluation for lower stakes testing contexts by Chapelle & Voss (2016) – that summaries of test-taker performance are accurate and relevant. Our warrant presumes that R-PLAT captures appropriate diagnostic descriptors of individual speaking ability needed for a strong learner fit quality of online language instruction. The pre-evaluation assumption underlying this warrant is that these diagnostic descriptors are indicative of target speaking ability levels used for placement into the respective levels of the course. Considering this assumption, we aim to answer the following research questions:

(1) Can raters' markings of diagnostic descriptors in R-PLAT serve as indicators of individual speaking ability?
(2) Can raters' descriptive comments in R-PLAT serve as indicators of individual speaking ability?

R-PLAT and the ITA Assessment Context

R-PLAT is a web-based assessment system that enables the delivery of the face-to-face institutional Oral English Certification Test (OECT), which is used for ITA certification and placement into level-based sections of the speaking course. It consists of two sections: an Oral Proficiency Interview (OPI) and a teaching simulation (TEACH). Following the ACTFL protocol, the OPI begins with an unrated "warm-up" introduction, which is followed by three impromptu questions and a role-play situation. The TEACH is intended to assess ITA candidates' ability to use English for teaching a topic in their field of study. Examinees present a mini-lecture on a topic in their discipline based on textbook materials they are provided. This mini-lecture is followed by a question-answer session, during which the raters ask questions about the presented content.

R-PLAT is designed as a rating platform for both OECT sections. It also integrates the scoring rubric and the OPI test items, which are provided as prompt sets developed based on ACTFL guidelines for eliciting a ratable sample. Each set contains proficiency-level-based impromptu questions on three different topics and different language functions (e.g., narrate, describe, compare, contrast, persuade, etc.), and a role-play task for four intended difficulty levels: advanced (230–300), intermediate-high (210–220), intermediate-mid (170–200), and intermediate-low (120–160). Individual students' performance is evaluated by two or three raters simultaneously; one of the raters acts as the interviewer in the OPI. All the raters use R-PLAT as follows:

(1) to assign a score for each response to each of the three OPI questions
(2) to assign a score for the performance on the OPI role-play
(3) to assign a total OPI score by averaging the four scores for each OPI prompt

(4) to assign a score for the TEACH performance
(5) to write descriptive comments about various aspects of language use in both OPI and TEACH.

Figure 8.1 demonstrates these affordances on the OPI page of R-PLAT.

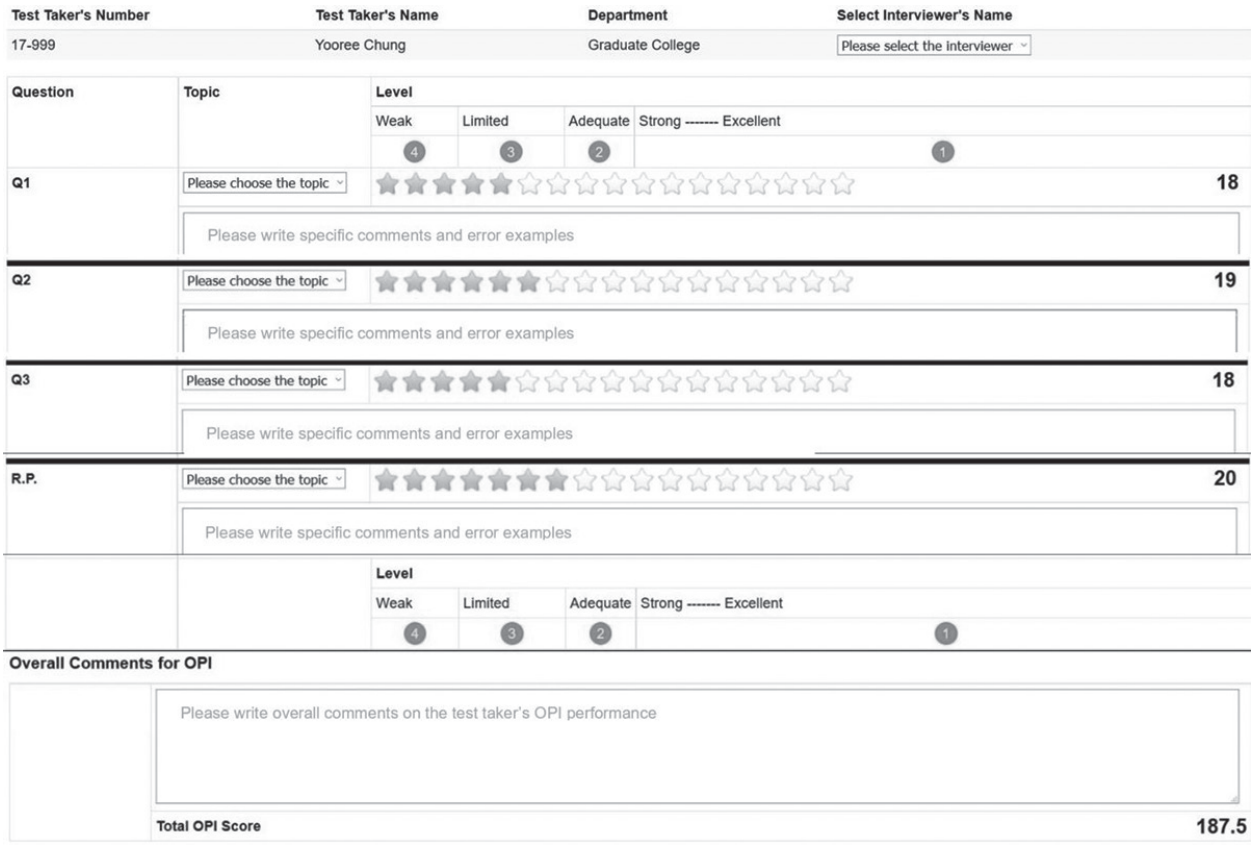

Figure 8.1. The OPI rating page in R-PLAT

Additionally, R-PLAT enables raters to mark up to 30 descriptors divided into seven diagnostic categories that are based on the OECT scoring criteria:

(1) Comprehensibility (ease of understanding, accent, and volume)
(2) Pronunciation (vowels, consonants, insertion, enunciation, reduction, intonation, rhythm, and word stress)
(3) Fluency (phrasing, choppiness, halting, false starts, pauses, incomplete utterances, and pace)
(4) Vocabulary (breadth of vocabulary, word choice, and expression)
(5) Grammar (grammatical complexity, word order, verbs, word form, singular or plural, pronouns, and articles)
(6) Pragmatics (interaction and compensation strategies)
(7) Listening.

As shown in Figure 8.2, each diagnostic descriptor within each category can be marked on a five-point scale.

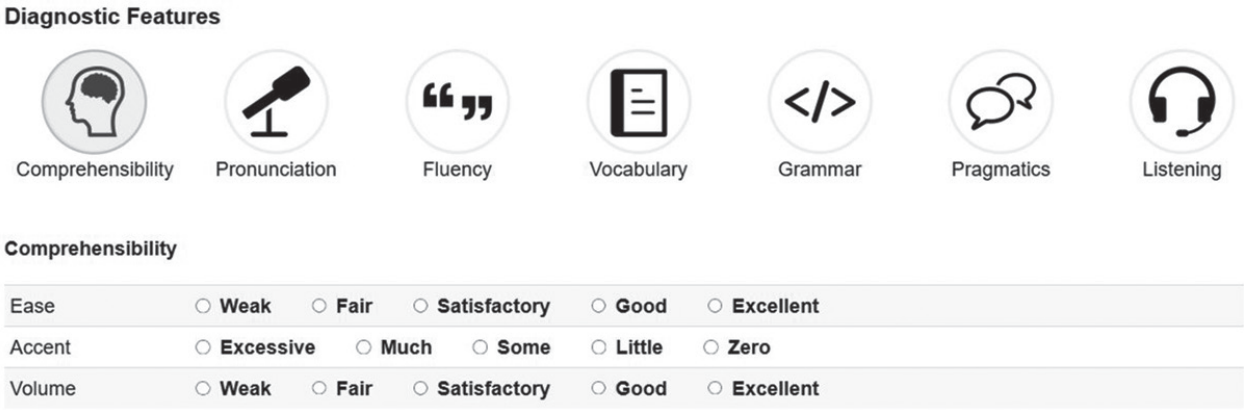

Figure 8.2. Diagnostic descriptors for Comprehensibility

With these affordances, R-PLAT facilitates language sample elicitation and evaluation processes. It also adds complementary functionality. For instance, the raters can access R-PLAT to verify their rating schedules. Test administrators can access the Administrator Portal in R-PLAT to adjudicate and finalize scores, then report the test results to students, their departments, and the instructors of ITA speaking courses into which students are placed. The system's database also contains demographic information including first language, graduate program, gender, etc. These speaking courses for ITAs have been offered face-to-face, but there is a pressing need for larger-scale and more learner-tailored online instruction. R-PLAT's affordances have thus been designed to capture multiple types of evidence indicative of prospective students' English speaking ability.

Participants

Eight OECT raters participated in this study. The raters had at least one-year rating experience and had participated in the assessment of ITAs prior to the implementation of R-PLAT. As is required before each test administration, the raters completed the so-called "brush-up" rater-training session. The rater-training included a review of test items and a tutorial for how to use R-PLAT. During the training sessions, raters practiced using R-PLAT for mock rating sessions with four video recordings and live testing of two ITAs. OECT data were collected from 53 prospective ITAs. These were graduate students in a wide variety of disciplines who were admitted to the university and considered for a teaching assistantship based on the following cut-off scores: TOEFL iBT 79, TOEFL PBT 550, IELTS 6.5, Pearson Test of English (PTE) 53.

Data and Data Analysis

A total of 146 OECT diagnostic ratings of the ITA participants were recorded during the administration of the test using R-PLAT. Of these, 50 ratings pertained to the advanced level, 39 to intermediate-high, and 57 to intermediate-mid. Intermediate-low is generally extremely rare, which is why this dataset did not contain this level. To examine the extent to which the diagnostic descriptors marked by the raters in R-PLAT indicated students' different speaking ability levels, the frequencies of 30 diagnostic descriptors marked on a five-point scale were collected from each rating. Henceforth, these will be referred to as diagnostic descriptor markings. In total, 2,524 markings for each proficiency level were analyzed. To compare the frequencies of these markings at each scale point (dependent variables) across the three

proficiency levels (independent variables), we ran Chi-Square tests, which are generally used to establish whether there is a relationship between two categorical variables (Larson-Hall, 2010). Next, seven Chi-Square tests were run separately to compare the markings across the three proficiency levels for each of the seven diagnostic indicators of speaking ability – comprehensibility, pronunciation, fluency, vocabulary, grammar, pragmatics, and listening.

In a similar vein, we examined whether the raters' descriptive comments can serve as diagnostic indicators of prospective ITA's speaking ability. The raters provided specific examples of patterns of language errors and/or appropriate language use as well as general impression comments about the test-takers' language performance on the OECT. The analysis of rater comments unfolded as follows. First, in the tradition of grounded theory (Glaser & Strauss, 1967), we identified themes in each rater's comments about individual test-takers' performance. The themes that emerged contained positive and negative assessments of language ability. The positive comments indicated strengths in language use with expressions such as "excellent," "strong," "good," etc. Negative comments highlighted weaknesses and examples of erroneous language use. The comments were further quantified using the metric of evaluative unit, defined as a segment (word, phrase, or clause) that expresses a rater's positive or negative assessment. For instance, the positive comment "*No effort to understand. Excellent enunciation and vocabulary.*" contains three evaluative units. In the negative comment, "*Some word stress issues. Lots of pausing and halting when nervous. Some sounds deleted.* ([*w*] *in wooden*)," four evaluative units are underlined. In total, 1,900 evaluative units were coded. Our reliability of coding rater comments into positive and negative evaluative units, measured by Cohen's kappa, was high ($k = 0.900$, $p = 0.000$).

To further establish if the diagnostic descriptors discriminate among the target speaking ability levels, the evaluative units were grouped by proficiency level: advanced (475 units), intermediate-high (523 units), and intermediate-mid (902 units). Additionally, the evaluative units were mapped onto each of the six criteria in the scoring rubric: functional competency (348 units), comprehensibility (150 units), pronunciation (526 units), fluency (424 units), vocabulary (167 units), and grammar (285 units). Cohen's kappa for coding into these evaluation criteria was also high ($k = 0.823$, $p = 0.000$). Descriptive statistics and Chi-Square tests were used to analyze differences in the frequencies of positive and negative evaluative units across the three proficiency levels, and for each of the six scoring criteria across the three proficiency levels. The dependency of the proficiency levels on the frequencies of positive and negative evaluative units was established by a Chi-Square test with a critical *p*-value of less than 0.05.

Results

Diagnostic Descriptor Markings

The first research question centered on raters' markings of the diagnostic descriptors enabled by R-PLAT. The analysis of raters' markings was conducted at two levels: first, 30 distinct descriptors were analyzed separately, and then the descriptors were grouped into seven higher-up diagnostic categories for analysis. Overall, the analysis provided evidence supporting our assumption that the diagnostic descriptors are indicative of target speaking ability levels. This can be inferred from the association between the markings of all diagnostic descriptors and the three proficiency levels depicted in Figure 8.3. The markings on the higher end of the scale (Good and Excellent) were more frequent for the advanced level, whereas the markings on the

lower end (Weak and Fair) were greater for intermediate-high and intermediate-mid.

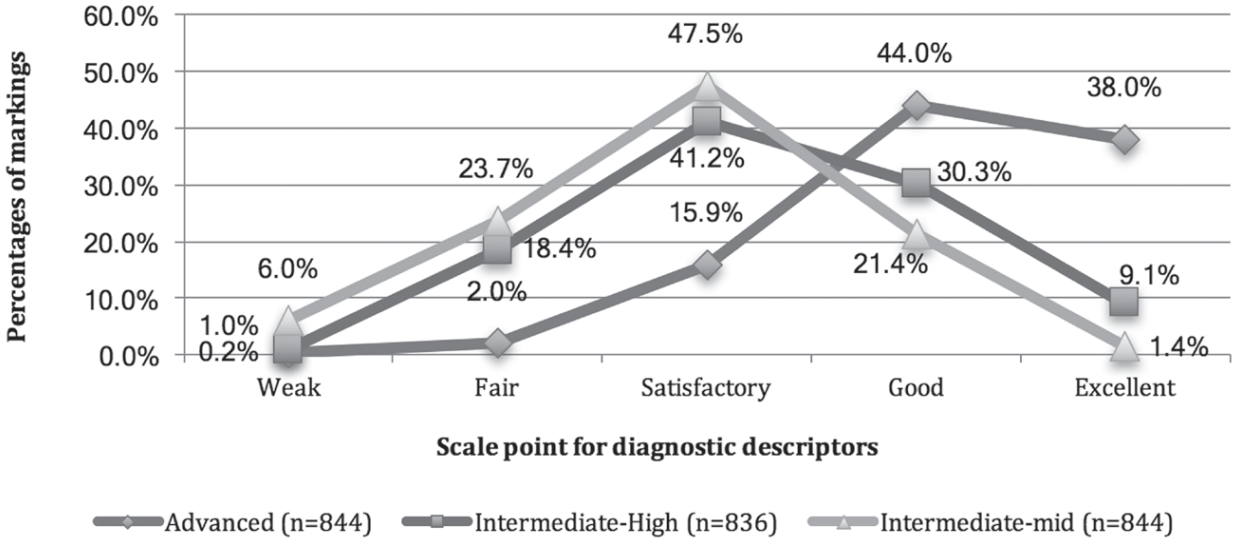

Figure 8.3. Distribution of the diagnostic descriptor markings on each scale across three proficiency levels

The Chi-Square test yielded significant differences in the frequencies of the markings at each scale point grouped by the three proficiency levels, X^2 (8, n markings = 2,524) = 830.92, $p < 0.01$. The more proficient students were given more markings higher on the scale compared to the less proficient students. These results suggest that the diagnostic descriptor markings support the proficiency ratings, especially distinguishing between the advanced and the two intermediate levels.

The results based on the markings grouped into diagnostic categories were similar. Those pertaining to five of the seven categories – Comprehensibility, Pronunciation, Fluency, Vocabulary, and Grammar – were consistently higher for the advanced level and lower for intermediate-mid and intermediate-high. The Chi-Square tests on these categories showed significant differences in the frequencies of the constituent descriptor markings at each scale point grouped by the three proficiency level ratings: Comprehensibility, X^2 (8, n = 340) = 83.99, $p < 0.01$; Pronunciation, X^2 (8, n = 707) = 185.95, $p < 0.01$; Fluency, X^2 (8, n = 610) = 232.02, $p < 0.01$; Vocabulary, X^2 (8, n = 178) = 116.53, $p < 0.01$; and Grammar, X^2 (8, n = 531) = 289.83, $p < 0.01$. The markings related to the other two categories – Pragmatics and Listening – exhibited somewhat different patterns. While the higher proficiency levels had more diagnostic descriptors marked as Excellent and Satisfactory and the vice versa, the percentages of descriptors marked as Good did not distinguish among the proficiency levels (Figures 8.4a and b). For Pragmatics, for instance, the percentages of the descriptors evaluated as Good clustered close together (advanced 42.6%, intermediate-high 42.6%, and intermediate-mid 40.0%). Nonetheless, these results are not surprising because pragmatic and listening abilities are not sub-traits of speaking proficiency. It is worth mentioning that the Chi-Square tests for the seven categories, including Pragmatics (X^2 (8, n = 158) = 43.26, $p < 0.01$) and Listening (X^2 (8, n = 73) = 34.41, $p < 0.01$), showed statistically significant differences. Thus, it can be inferred that the diagnostic descriptor markings for the seven categories were distributed at each scale point reflective of different proficiency levels.

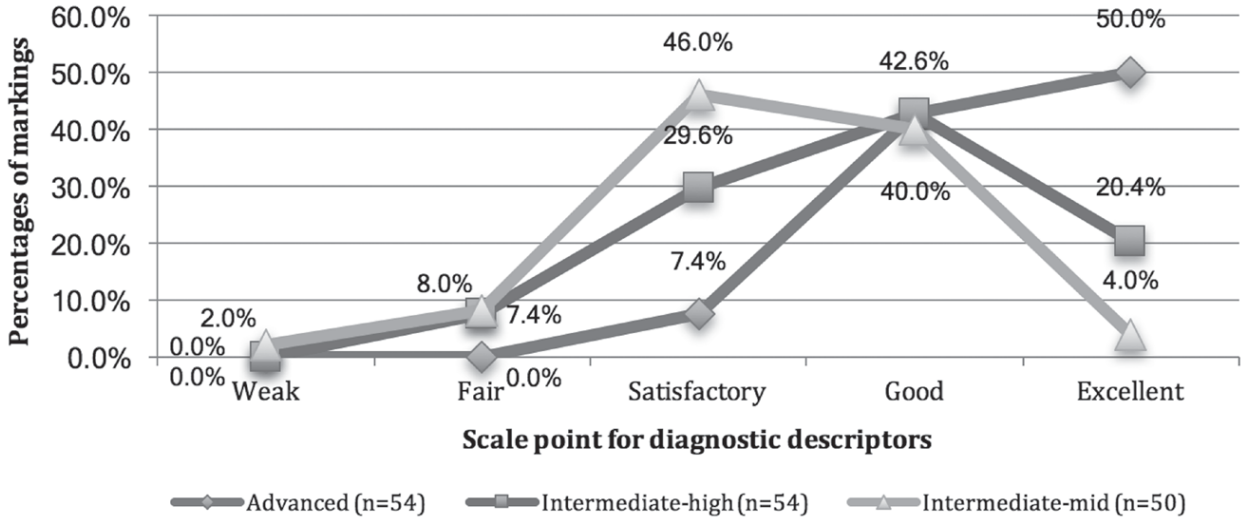

Figure 8.4a. Distribution of diagnostic descriptor markings for Pragmatics on each scale across three proficiency levels

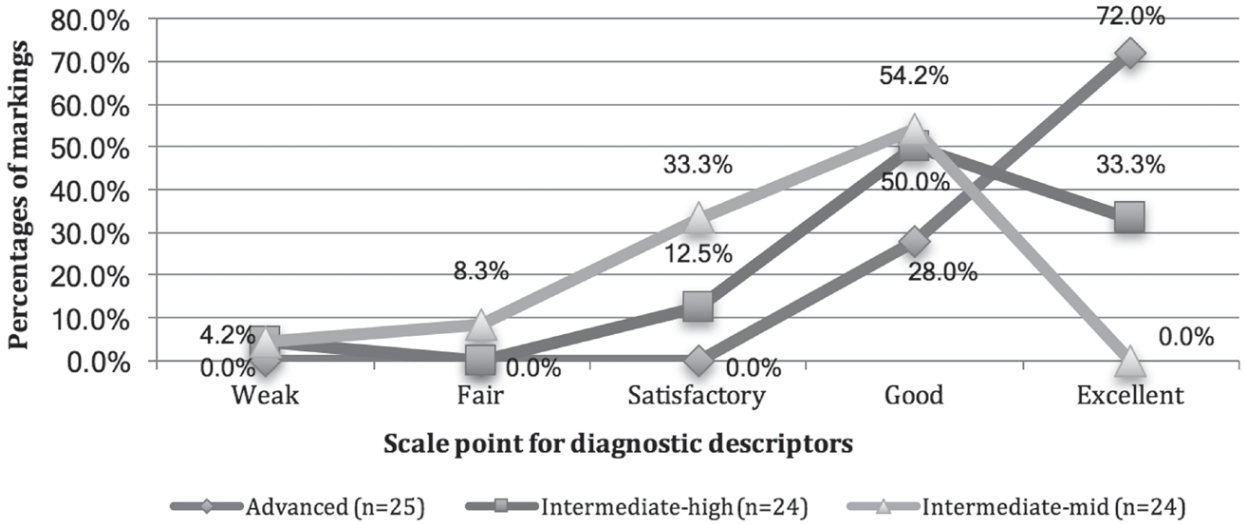

Figure 8.4b. Distribution of diagnostic descriptor markings for Listening on each scale across three proficiency levels

Descriptive Comments

To answer the second research question, we analyzed the diagnostic information provided by raters in the form of open-ended comments enabled by R-PLAT. Here, too, our assumption appeared to be supported. The analysis of positive and negative evaluative units exhibited clear associative patterns between the type of evaluative units and the three proficiency levels. The percentage of positive evaluative units was higher for the advanced level (61.0%) and much lower for intermediate-high (18.5%) and intermediate-mid (13.4%); the negative evaluative units indicated the opposite (Figure 8.5). Chi-Square results showed significant differences in the

frequencies of positive and negative evaluative units across the three proficiency levels, X^2 (2, n = 1,900) = 386.50, $p < 0.01$. Overall, the comments were reflective of students' different proficiency levels and can thus be considered as diagnostic indicators of speaking ability.

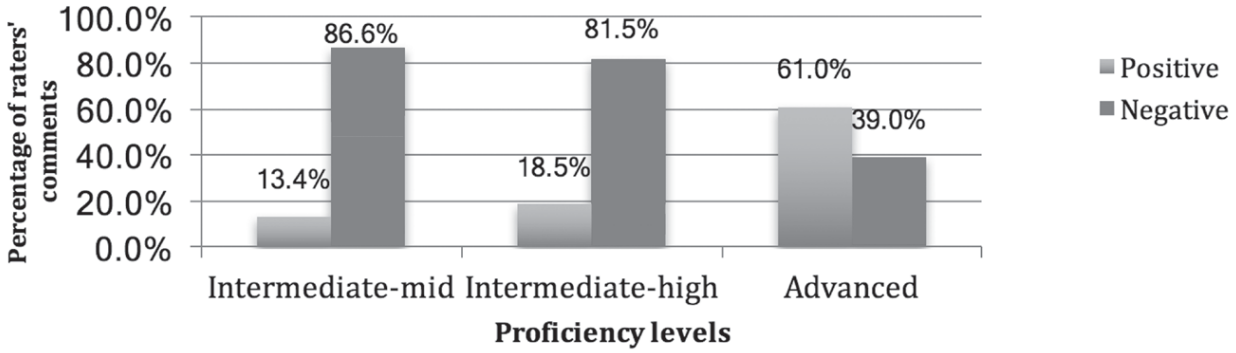

Figure 8.5. Overall distribution of positive and negative evaluative units across proficiency levels

In a parallel strand of analysis, the positive and negative evaluative units were grouped based on the OECT scoring criteria: functional competency, comprehensibility, fluency, vocabulary, pronunciation, and grammar; and then compared across proficiency levels. Descriptive comments related to the first four criteria exhibited patterns similar to the one in Figure 8.5. For grammar and pronunciation, the number of negative evaluative units exceeded that of positive units regardless of proficiency level. The percentages for negative units were much greater than those for positive evaluative units (e.g., pronunciation in Figure 8.6). This is likely due to the fact that the raters pay a lot of attention to language errors, recording which helps them determine if those appear to be error patterns or mistakes. It is also possible that the prompt in R-PLAT's comment box ("Please write specific comments and error examples") encourages them to focus on errors. All Chi-Square tests for each scoring criterion showed significant differences in the frequencies of positive and negative evaluative units across levels, confirming the association of the two types of evaluative units with the three proficiency levels.

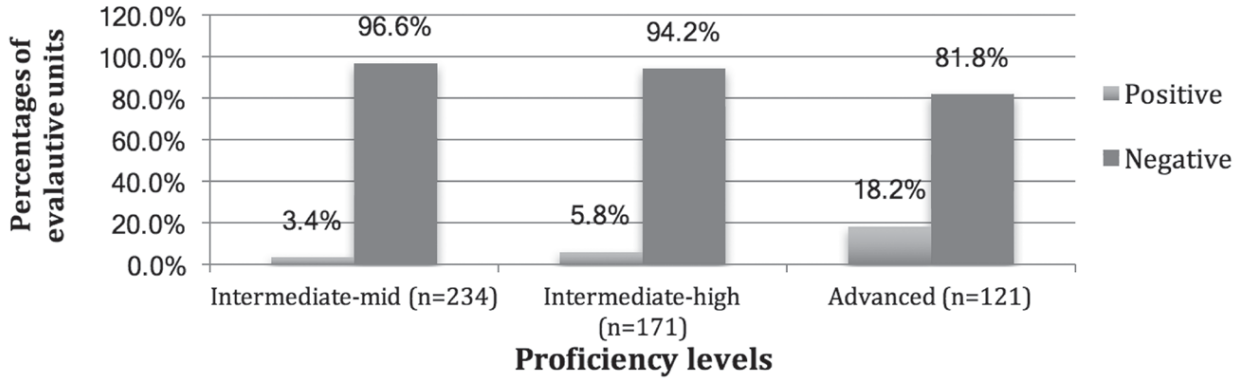

Figure 8.6. Distribution of positive and negative evaluative units across proficiency levels for pronunciation

Discussion

The results of this first effort towards supporting assumptions about the intended use of R-PLAT will be used to develop a principled approach to designing on-line ITA instruction with a learner fit quality substantiated by reliable diagnostic information. In the face-to-face course, individualizing teaching from the very beginning has been beyond the bounds of possibility. R-PLAT's affordances, however, will greatly facilitate transition to the online mode and will allow for a fuller integration of proficiency assessment in online course design.

ITA program administration will use R-PLAT-recorded data to develop a formalized, multifaceted spec model of diagnostic descriptors, containing clearly defined characteristics and representative of both proficiency level and first language. The diagnostic spec model will enable tailoring the curriculum to more specific needs. Most importantly, a judgmental analysis will be conducted to determine whether the desired linguistic features detected through R-PLAT would appear in the course materials. Once level-based diagnostics-to-materials mapping is accomplished, multiple types of enhancement will be applied to the desired features. The model will also become an essential resource for teacher and rater training. Additionally, the OECT performance-level descriptors will be revised in view of the new spec model to improve the level of detail often sought not only by teachers and students, but also by raters, course and test coordinators, and students' graduate programs.

Instructors, successively, will be provided with diagnostic information derived from R-PLAT about the students placed in the section of the course they will be teaching, which they will use to develop learning and practice plans for individual students given their proficiency level and language background. To capacitate teachers to tailor the syllabus to group-specific needs, diagnostic group-level reports will also be supplied. This will help teachers determine the appropriateness of the level of linguistic difficulty and adapt the materials and tasks in order to focus on the desired linguistic forms. For this purpose, the 30 diagnostic descriptors for both segmental (e.g., vowels, consonants) and suprasegmental (e.g., intonation, stress) features will be particularly useful. In the future, R-PLAT could be customized further to enable diagnostically-driven formative feedback by the teacher as well as by peers.

The implications of this study extend to curriculum-based assessment in online language education as well. Language tests are generally categorized based on the type of decisions they are used to inform. OECT serves the dual purpose of assessing ITAs' level of speaking ability and to decide which section of the course they should be placed in. While facilitating the former, the value of R-PLAT is in its unique capability to catalog diagnostic evidence in order to inform curriculum design and the learner fit quality of online teaching. This purpose overlaps with diagnostic assessment where the goal is to exert positive change in student learning; yet, it is slightly different in that the focus is on curriculum and course design. Considering this primary focus on pedagogy, it seems to us that a new form of curriculum-based assessment could emerge – assessment for language instruction (ALI). Despite its focus on learning tasks, ALI would be distinct from learning-oriented language assessment. The latter "is inherently interactive" and entails learner-involved assessment (i.e., self and peer evaluation) and learning-focused feedback (i.e., interlocutor scaffolding, immediate feedback, and focus on feed-forward), which are essential during teaching and learning but immaterial for learner diagnostics-based pre-evaluation of instruction.

On a broader scale, such an innovation in assessment practices seems to be particularly called for, given the proliferation of web-based instructional environments. As Chapelle and Voss (2016: 120, 121) argue, "[a]n innovative agenda for language assessment extends beyond the

goal of making more efficient tests to expanding the uses of assessment and their usefulness"; and technology is "ideally suited to play a role in this vision because of their capacity for individual treatment of test takers as learners," who "should actually be given opportunities to learn from both the process and the results of test taking." Apart from instantiating positive changes in learning, ALI would impact teachers, who would no longer be pressed to adapt haphazardly to their subjective interpretations based on the results of diagnostic tests, which may or may not be developed. Instead, they would be guided by reliable diagnostics, perhaps from several raters, as it is the case with R-PLAT.

Conclusion

This study investigated the potential of R-PLAT to capture diagnostic evidence from a speaking test. The data consisted of diagnostic descriptor markings and descriptive comments about individual students' oral proficiency provided by raters via R-PLAT. Our judgmental assumption that these types of diagnostic evidence are indicative of speaking ability was substantiated by the associations between the diagnostic descriptors and the target proficiency levels, which are used for placement into respective levels of the speaking course for ITAs. Continuing this research agenda, each aspect of online course development informed by this study will be investigated as warrants and assumptions for a range of inferences in the validity argument for R-PLAT. Although there is much yet to do, we have molded the first building block for leveraging technology to collect multiple types of diagnostic evidence for assessment for online language instruction.

About the Authors

Hyejin Yang is a Lecturer at Sookmyung Women's University and Yonsei University, Korea. Her research interests include language assessment, computer-assisted language learning, and L2 speaking and writing.

Elena Cotos is an Assistant Professor in the Applied Linguistics Program at Iowa State University. She is also the Director of the Center for Communication Excellence of the Graduate College at Iowa State University. She investigates computer-assisted language learning and assessment, written and spoken genres, and automated writing evaluation.

References

ACTFL (American Council on the Teaching of Foreign Languages), (2012). *ACTFL Oral Proficiency Interview Familiarization Manual*. Retrieved December 12, 2016 from http://www.languagetesting.com/wp-content/uploads/2013/05/ACTFL-OPI-Familiarization-Manual1.pdf.

Bachman, L. F., & Palmer, A. S. (2010). *Language Assessment in Practice: Developing Language Assessments and Justifying Their Use in the Real World*. Oxford: Oxford University Press.

Blake, R., Wilson, N., Cetto, M., & Pardo Ballester, C. (2008). Measuring oral proficiency in distance, face-to-face, and blended classrooms. *Language Learning & Technology*, 12(3), 114–127. Retrieved from http://llt.msu.edu/issues/june2016/blake.pdf

Carr, N. T. (2011). *Designing and Analyzing Language Tests*. New York, NY: Oxford University Press. doi: 10.1017/s0272263112000800

Chapelle, C. (2001). *Computer Applications in Second Language Acquisition: Foundations for Teaching, Testing, and Research*. Cambridge: Cambridge University Press.

Chapelle, C. (2009). The relationship between second language acquisition theory and computer-assisted language learning. *Modern Language Journal*, 93(S1), 741–753. doi:10.1111/j.1540-4781.2009.00970.x

Chapelle, C. A., Chung, Y-R., & Xu, J. (eds.). (2008). *Towards Adaptive CALL: Natural Language Processing for Diagnostic Language Assessment.* Ames, IA: Iowa State University.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (eds.). (2008). *Building a Validity Argument for the Test of English as a Foreign Language*^{TM}. New York:Routledge. doi:10.4324/9780203937891

Chapelle, C. A., & Voss, E. (2016). 20 years of technology and language assessment in language learning & technology. *Language Learning & Technology*, 20(2), 116–128. Retrieved from http://llt.msu.edu/issues/june2016/chapellevoss.pdf

Doughty, C. J., & Long, M. H. (2003). Optimal psycholinguistic environments for distance foreign language learning. *Language Learning & Technology*, 7(3), 50–80. Retrieved from http://llt.msu.edu/vol7num3/pdf/doughty.pdf

Farnsworth, T. L. (2013). An investigation into the validity of the TOEFL iBT speaking test for international teaching assistant certification. *Language Assessment Quarterly*, 10, 274–291. doi:10.1080/15434303.2013.769548

Fox, J. (2009). Moderating top-down policy impact and supporting EAP curricular renewal: Exploring the potential of diagnostic assessment. *Journal of English for Academic Purposes*, 8, 26-42. doi:10.1016/j.jeap.2008.12.004

Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of Grounded Theory*. New York: Aldine.

Hubbard, P. (1988). An integrated framework for CALL courseware evaluation. *CALICO Journal*, 6(2), 51–72. doi: 10.1558/cj.v6i2.51-72

Hubbard, P. (2006). Evaluating CALL software. In L. Ducate & N. Arnolds (eds.), *Calling on Call: From Theory and Research to New Directions in Foreign Language Teaching* (pp. 313–334). San Marcos, TX: CALICO.

Jamieson, J., Chapelle, C., & Preiss, S. (2005). CALL Evaluation by developers, a teacher, and students. *CALICO Journal*, 23(1), 93–138. doi:10.1558/cj.v23i1.93-138

Jang, E. (2012). Diagnostic assessment in language classrooms. In G. Fulcher & F. Davidson (Eds.), *The Routledge Handbook of Language Testing* (pp. 120–133). New York: Routledge. doi:10.4324/9780203181287

Kane, M. T. (2006). Validation. In R. L. Brennan (ed.), *Educational Measurement* (4th ed.) (pp. 17–64). Westport, CT: American Council on Education.

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, 23:2, 198–211. doi:10.1080/0969594x.2015.1060192

Larson-Hall, J. (2010). *A Guide to Doing Statistics in Second Language Research Using SPSS*. New York: Routledge.

Levy, M. & Kennedy, C. (2004). A task-cycling pedagogy using stimulated reflection and audio-conferencing in foreign language learning. *Language Learning & Technology*, 8(2), 50–69. Retrieved from http://llt.msu.edu/vol8num2/pdf/levy.pdf

Lim, H., Kim, H., Behney, J., Reed, D., Ohlrogge, A., & Lee, J. E. (2012, March). Validating the use of iBT Speaking scores for ITA screening. Paper presented at the TESOL Annual Convention and Exhibit, Philadelphia, PA.

Nichols, P. D., Meyers, J. L. & Burling, K. S. (2009). A framework for evaluating and planning assessments intended to improve student achievement. *Educational Measurement: Issues and Practice*, 28(3), 14–23. doi:10.1111/j.1745-3992.2009.00150.x

Volle, L. M. (2005). Analyzing oral skills in voice e-mail and online interviews. *Language Learning & Technology*, 9(3), 146–163.Retrieved from http://llt.msu.edu/vol9num3/pdf/volle.pdf

Wang, H., Choi, I., Schmidgall, J., & Bachman, L. F. (2012). Review of Pearson test of English academic: Building an assessment use argument. *Language Testing*, 29(4), 603–619. doi:10.1177/0265532212448619

Wylie, E. C., & Tannenbam, R. J. (2006). *TOEFL® Academic Speaking Test: Setting a Cut Score for International Teaching Assistants*. Research Memorandum (RM-06-01). Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Tests/TOEFL/pdf/ngt_itastandards.pdf

Xi, X. (2007). Validating TOEFL® iBT Speaking and setting score requirements for ITA screening. *Language Assessment Quarterly*, 4(3), 318–351. doi:10.1080/15434300701462796

Xu, H. & Morris, L. V. (2007). Collaborative course development for online courses. *Innovative Higher Education*, 32, 35–47.doi:10.1007/s10755-006-9033-5