Becker *et al. Asian-Pacific Journal of Second and Foreign Language Education* (2017) 2:18
DOI 10.1186/s40862-017-0041-z

Asian-Pacific Journal of Second and Foreign Language Education

ORIGINAL ARTICLE Open Access

Balancing practicality and construct representativeness for IEP speaking tests

Anthony Becker[1], Sawako Matsugu[2] and Mansoor Al-Surmi[3*]

* Correspondence: alsurmi@qu.edu.qa
[3]Department of English, Foundation Program, Qatar University, P.O. Box 2713, Doha, Qatar
Full list of author information is available at the end of the article

Abstract

Intensive English programs (IEPs) strive to make certain that international students have sufficient levels of speaking ability, which is typically assessed through a combination of different tasks. One drawback of including multiple tasks is that the development, administration, and scoring might not be practical. Therefore, it is important to investigate how well the tasks account for examinees' speaking ability, as using fewer tasks could help in minimizing resources. Using quantitative methods of analysis, this study evaluates how well four types of speaking tasks on proficiency and achievement tests account for students' speaking ability in an IEP. The findings indicate that several tasks uniquely contribute to the speaking construct. This study has implications for the importance of balancing practicality with construct representativeness and presents a model of how IEPs might approach this issue.

Keywords: Construct representativeness, Practicality, Speaking, Test use, IEP

Introduction

Speaking ability in north American universities

The ability to speak well in a second language is important for English language learners (ELLs) studying at an English-medium university. Speaking, as a social and situation-based activity, is essential for ELLs' basic interpersonal communication skills (BICS), or the language skills needed in social situations, and their cognitive academic language proficiency (CALP), or the language needed for formal academic learning (for an explanation of BICS/CALP, see Cummins, 2008). As Douglas (1997) indicates, ELLs require strong speaking skills to perform day-to-day language functions (e.g., ordering food, buying textbooks), as well as to discuss content in the classroom and promote understanding within the academic domain. For that reason, ELLs must have pre-requisite speaking skills in English, if they are to successfully engage in the wide variety of tasks requiring speaking in an academic context.

Richards (2008) contends that ELLs must demonstrate control of three primary speech activities, all of which are commonly found in academic settings at North American universities: 1) talk as interaction, 2) talk as transaction, and 3) talk as performance. *Talk as interaction* refers to everyday conversational discourse that primarily serves a social function. These exchanges require ELLs to exchange greetings, engage in small talk (e.g., chatting about their weekend), and share personal experiences. *Talk as transaction* refers to situations where ELLs must focus on the clarity and accuracy

of their spoken message; such exchanges typically include focused speech activities, such as classroom group discussions, office-hour visits, buying something from the store, and ordering food from a restaurant. *Talk as performance* refers to situations where ELLs are required to transmit information before an audience, including doing classroom presentations, delivering speeches, making announcements, or conducting a debate. Given the variety of speaking situations that ELLs will encounter, many North American universities have created intensive English programs (IEPs), which offer instruction to prepare ELLs for the language (and academic) demands required in mainstream university classes.

IEPs are university-based programs in which students participate in a large number of accelerated English language courses. These programs, which primarily provide English for academic purposes, are sheltered learning environments whereby students receive instruction to improve their general and academic English skills; they are designed to optimize the learning time in roughly four to five hours a day (Nasri & Shokrpour, 2012; Stoynoff, 2002). The goal of most IEPs is to provide students with the language skills and learning strategies they need to develop communicative competence in English, for the purpose of helping them to succeed in mainstream university classes (Hillyard et al., 2007). While all four language skills (i.e., listening, reading, speaking, and writing) are typically taught and assessed at IEPs, there is usually a strong emphasis placed on developing oral proficiency, as it has been found to be closely linked to academic success (North, 2000; Powers, 2010). Therefore, by the end of IEP studies, ELLs are expected to have a high level of oral proficiency. With this goal in mind, IEPs strive to assess such development at multiple levels of the program.

Assessment of speaking skills

In developing tasks for the assessment of speaking skills, there needs to be a theoretical model that serves as the basis for such tasks. Bachman and Palmer's (1996, 2010) model of language ability (described below) is the most widely-recognized communicative model in language assessment. Within their framework, language use is considered as the interaction between language users and the context in which they belong to. Language use is influenced by characteristics of individuals, including personal characteristics, topical knowledge, affective schemata, and by language ability, which includes language knowledge and strategic competence. Language knowledge refers to organizational knowledge and pragmatic knowledge, whereas strategic competence refers to goal-setting, assessment, and planning (see Bachman & Palmer, 1996, for a detailed overview of their framework). The utility of the framework is that the knowledge and strategies in the language ability segment of the model can be applied to various speaking situations, as tasks can be considered language use situations (Luoma, 2004).

Speaking tasks can be categorized in several different ways. One way is to consider the number of examinees involved in performing a task. This can be thought of in terms of the three task types commonly used to elicit speech samples: individual, pair, and group tasks (Luoma, 2004). Examples of individual tasks include interviews (one interviewer and one interviewee) and independent speaking tasks, in which examinees are asked to state their opinions about a given topic or to respond to a situation. Narrating stories or situations also falls under this individual task category. Furthermore,

individual tasks can incorporate the use of multiple language skills (e.g., reading and/or listening), such as the integrated speaking tasks used in the TOEFL iBT. Pair tasks can be conducted in the form of an interview (one interviewer, two examinees in the pair), or when two examinees are asked to make a joint decision or provide recommendations for a given issue. Paired-speaking tasks require peer-to-peer interaction that results in an extended exchange of information (Luoma, 2004). Finally, group tasks involve three or more examinees and typically share the same communicative goals as paired-speaking tasks. For assessment purposes, group tasks are not as common in classroom settings as independent and paired-speaking tasks because of the challenges that they present (for group work challenges, see Davies, 2009); however, group tasks are often used in conjunction with those tasks in the format of group discussions after individual presentations (Luoma, 2004).

At the IEP described in the present study, independent speaking, integrated speaking, picture narrative, and paired-speaking tasks have been traditionally used in the program's achievement and proficiency tests. Each of these task types helps to provide evidence of our ELLs speaking abilities, as each of the speaking tasks offers some unique (albeit sometimes overlapping) information. For instance, independent speaking tasks result in monologic spoken discourse on a variety of topics by incorporating speakers' own knowledge and/or experience through a variety of rhetorical/pragmatic functions. This task type mainly measures intelligibility and fluency (Enright et al., 2008). The integrated task, on the other hand, requires examinees to use information from a range of academic texts (for reading) or consultations and student interactions (for listening) as the basis for creating monologic discourse using rhetorical/pragmatic functions. This task type measures intelligibility, fluency, content, coherence, and organization (Enright et al., 2008). The narrative task elicits examinees' ability to state a sequence of events through monologic discourse. With narrative tasks, examinees are often asked to set the scene and identify the main characters and events in a chronological order. With respect to paired-speaking tasks, they are considered open-ended, dialogic tasks. Speakers are generally allowed to choose the direction of the speech and to what extent they need to meet the task requirements. According to French (2003), the paired-speaking task can lead examinees to produce a wider range of language functions compared to those elicited by interview tasks. The role that each partner plays in the paired-speaking task, combined with the range of language functions utilized, may contribute to more inclusive speech samples from examinees (Skehan, 2001). Overall, each of these tasks contributes to the construct of speaking ability envisaged at the IEP discussed in this study.

Defining the construct of speaking

In second language assessment, a comprehensive definition for speaking ability does not currently exist. This is due to the fact that the nature of speaking is quite sophisticated and multifaceted. Speaking is so intertwined with our everyday use of language that it becomes difficult to create a concise, yet widespread, definition of speaking (Thornbury & Slade, 2006). While various frameworks of speaking ability have been proposed for specific situations (see Luoma, 2004), second language researchers have generally relied on defining speaking by its features, functions, and conditions.

According to Nazara (2011), speaking entails three areas of knowledge and/or use. The first area involves the mechanical (i.e., linguistic) features of language, which refer to the grammar, pronunciation, and vocabulary demonstrated in spoken discourse. These features enable a speaker to implement the most fitting words in the correct sequence, with the appropriate pronunciation. The second area entails the functions of spoken language, which focus on the transactional and interactional uses of speech. These functions allow a speaker to exchange information and to understand when clarification is needed (e.g., when miscommunication occurs in a transaction). The third area concerns the socio-cultural norms associated with different speech situations (e.g., turn-taking, rate of speech, and roles of participants). Knowledge of these norms permits a speaker to understand the conversational situation, to realize who is being spoken to, and to comprehend the purpose of the speech act (Nazara, 2011). Taken together, these three characteristics of speaking serve to describe how it can be defined. While these characteristics can be applied to the speaking ability in general, their use here is limited to defining the speaking ability in an academic context.

For the present study, the construct of academic speaking has been defined as follows: academic speaking is the ability, when communicating with others, to use oral language comprehensibly and appropriately to acquire, transit, and demonstrate knowledge (Butler et al., 2000; Fulcher, 2003; Luoma, 2004; Pearlman, 2008). In the IEP where the study was situated, efforts were made to cover the defined construct here through different types of speaking tasks at different levels of English language study.

Practicality

When assessing speaking skills, the practicality of the tasks being used influences the extent to which the speaking construct can be represented, since the types of speaking tasks that can actually be administered depend largely on the resources that are available. The issue of practicality when assessing speaking is especially important for IEPs, since many speaking tasks used for assessment purposes [in IEPs] tend to be performance-based, in order to reflect the types of speaking skills that ELLs need to demonstrate in the university setting (Gan, 2012). While it may be desirable to represent the speaking construct as broadly as possible, given the fact that ELLs will encounter numerous speaking situations in the university setting, there must be a balance between what is desirable and what is practical. In this respect, a general plan is needed for determining the relationship between the resources that are available and what is required in the design, development, and use of a speaking assessment.

Bachman and Palmer (1996) provide a helpful framework for determining the practicality of an assessment, which can be applied to situations of assessing speaking at IEPs. As shown in Table 1, they indicate that assessment development and use should include the consideration of human and material resources, as well as time. In addition, we argue that considerations for monetary resources are equally as important, since the availability of funding greatly influences the human and material resources that are possible for developing and administering an assessment.

Table 1 Considerations for determining the practicality of an assessment

Consideration	Description
Human Resources	Availability of people (e.g., assessment task creators, scores or raters, test administrators, and clerical support)
Material Resources	Space (e.g., rooms for assessment, development, administration) Equipment (e.g., tape, video, DVD recorders, computers) Materials (e.g., paper, pictures, library resources, computer software)
Time	Development time (e.g., time from the beginning of the assessment development process to the recording of scores)
Monetary Resources	Time for specific tasks (e.g., designing, creating, administering, scoring, analyzing) Budget (e.g., personnel, production, and equipment costs)

Taking the issue of practicality into consideration, the current study aims to determine the extent to which speaking tasks used in a proficiency test (i.e., picture-narrative, independent, and integrated tasks) and speaking tasks used in an achievement test (i.e., paired summary, and integrated tasks) represent the speaking construct for English language tests used at an IEP in a North American university. In addition, the study aims to help determine whether using fewer speaking tasks is more practical to represent the speaking construct. The research questions for the present study are as follows:

1. To what degree do the speaking tasks overlap?
2. What practical resources are associated with two versus three tasks?

Methods

Setting

The IEP in this study is located at a southwestern US university and was established to prepare international students for academic success in mainstream university classes by improving their English proficiency and academic skills. At the time of the study, the majority of students were from China and Saudi Arabia. There were also students from Korea, Japan, United Arab Emirates, Kuwait, and multiple other southeast Asian countries. Most students were in their late teens or early twenties.

All in-coming international students were required to take a placement test at the IEP unless they submitted standardized English test scores, such as the Test of English as a Foreign Language (TOEFL) or the International English Language Testing System (IELTS), that met the university's direct admissions criteria. Depending on their performance on the placement test, students could be unconditionally admitted to the university, or placed into an appropriate level at the IEP for English instruction before advancing to mainstream university classes.

There were four levels in the IEP when the study was conducted: Pre-academic, Level 1, Level 2, and Bridge. Pre-academic level was equivalent to scores of 400 or below on the paper version of TOEFL. Level 1 was 400 to 449, 450 to 489 for Level 2, and 490–524 for Bridge. Students received 13 to 27 h of instruction (per week) depending on their assigned level. Students at all levels were required to take six hours of Listening/Speaking, and another six hours of Reading/Writing for Pre-academic, Level 1, and Level 2, and Reading/Vocabulary for Bridge level. Students in the Bridge level were also allowed to take six credits of specified courses (i.e., English composition for non-native speakers of English) at the university.

Testing at the IEP

Proficiency tests (placement and exit tests)

Proficiency tests were administered twice a semester: at the beginning and end of the same semester. They were developed by assessment coordinators working at the IEP. The tests assessed four language skills (i.e., listening, speaking, reading, and writing). Each test took approximately three hours to complete. All tests were scored by trained staff at the IEP. In order to assess students' speaking abilities, picture narrative, independent, and integrated tasks were given as a part of the proficiency test. Students were given approximately one minute to prepare their speech and one minute to respond to their tasks. See Table 2 for a description of the speaking tasks in the proficiency test.

Achievement tests

Three achievement tests were administered for the Listening/Speaking and Reading/Writing courses during weeks 5, 10, and 15 in the semester during which the study was conducted. Final grades in those courses were determined based on the results of these achievement tests, which accounted for 40% of the final grade, in addition to other assignments, such as homework and projects. The first test was given a weight of 10%, and 15% for the latter two achievement tests, respectively. These tests, which were developed by course instructors and the assessment coordinators, were based on the textbooks that were used in the courses. The speaking tasks given as a part of the achievement tests were paired, summary, and integrated tasks in Levels 2 and Bridge, both of which were included in the analysis for the present study. See Tables 3 and 4 for descriptions of the speaking tasks included in the two achievement level-tests.

Data collection

Speaking scores were collected for a total of 133 ELL students from five speaking task-types that were part of the achievement and proficiency tests (described below). Specifically, the speaking scores for 47 students were collected for two classroom-based achievement tests (Level 2 and Bridge) that were administered during weeks 10 and 15 of the semester; the speaking scores for 86 students were collected for one proficiency test (used to make placement decisions) that was administered at the end of the same semester. All students were enrolled in the IEP when the data was collected for this study. The average age among students was approximately 20 years old.

The achievement tests of students from Level 2 and Bridge were chosen for data collection, as these tests included three different speaking tasks, versus other levels at the

Table 2 Speaking tasks in the proficiency test

Tasks	Preparation time	Response time	Prompt
Narrative	45 s.	1 min.	Look at the six pictures carefully. The pictures are already ordered. Create a story based on the pictures. Be sure to: 1) connect each picture in your story, and 2) give a lot of details.
Independent (opinion)	45 s.	1 min.	Some people like to stick to activities they know they can do well. Others like to try new things and take risks. Which do you prefer? Why?
Integrated (listen & give opinion)	45 s.	1 min.	Summarize Tom Mortenson and Linda Hallman's views of gender in education. Do either of these views describe your country's current educational situation?

Table 3 Speaking tasks in an achievement test (Bridge: Low Advanced)

Tasks	Preparation time	Response time	Prompt
Paired	1 min.	3 min.	Decide whether a patient should be told the extent of their illness.
Summary (listen & summarize)	1 min.	1 min.	Listen to the discussion. After listening, prepare an oral summary of the main ideas and important details.
Integrated (listen, summarize, give opinion)	1 min.	1 min.	You just listened to a doctor discuss some ideas related to shyness. Would you agree or disagree with the doctor's ideas about shyness? Explain your answer using details and examples.

IEP that only included two speaking tasks. The speaking tasks in the achievement tests were given during regular class hours as part of the test for Listening/Speaking courses in both levels. For the paired-speaking task, students were paired by their instructor in advance and students recorded their responses for each task using digital recorders. The students' instructor proctored all of the task administrations and all speech samples were scored by trained teacher-raters at the IEP.

The speaking tasks in the proficiency test were given as part of the placement test following writing, listening, and reading sections. These tasks were administered in person in a separate room and students' responses were recorded with a digital recorder. All scoring for the speaking tasks in the proficiency test was completed by trained teacher-raters at the IEP.

Analysis

To answer the first research question, descriptive statistics (means and standard deviations) were computed. Then, Pearson product-moment correlation coefficients were computed among scores in tasks for each test to investigate how strong the relationships were among the tasks and the construct of speaking. Large correlations ($\geq \pm .70$) among tasks were seen to provide evidence of convergence, which would indicate the assessment of the same construct. Low correlations ($\leq \pm .39$) among tasks did not provide evidence of convergence, which would imply the assessment of a different construct. See Grimm (1993) for a discussion of correlations and their relation to providing evidence of convergent validity.

In terms of the second research question, the costs in terms of human resources, materials, time, and money that were required in developing, administering, and scoring

Table 4 Speaking tasks in an achievement test (Level 2: Intermediate)

Tasks	Preparation time	Response time	Prompt
Paired	1 min.	3 min.	Students each have information on one applicant. They decide which applicant to hire.
Summary (listen & summarize)	1 min.	1 min.	Listen to the Radio show. Summarize the main ideas and important details.
Integrated (listen, summarize, give opinion)	1 min.	1 min.	Professor John Gibson and Mr. Daniel Tucker are discussing television news. Do you agree with their opinion regarding the problems and benefits of daily news on TV? Provide supporting details.

the tasks in the achievement and placement tests were calculated. These figures were compared to determine the cost effectiveness of the development, administration, and scoring of the speaking tasks used in the IEP achievement and proficiency tests. In the following section, results based on the statistical analyses and test cost calculations are reported. In addition, our decisions for how to balance the construct representativeness and practicality of the speaking tests are discussed.

Results

Using test scores of 133 ELLs, the study investigates whether the speaking construct in English language tests is similarly represented across speaking tasks used in placement and achievement tests used at an IEP. The results are presented as they relate to the two research questions for the study.

The first research question examines the extent to which the speaking tasks utilized in each test overlap. To address this question Pearson's *r* correlation coefficients among the speaking test tasks were calculated. As indicated earlier, if large correlation coefficients were obtained (i.e., ≥ ± .70), this would provide evidence that the tasks were assessing the same construct. A total of 5 points was possible for each task, with the mean score for the narrative task being the highest ($M = 4.26$, $SD = 1.84$), followed by the independent task ($M = 4.05$, $SD = 2.11$) and the integrated task ($M = 3.09$, $SD = 1.95$). Overall, there were large correlations among the tasks of the proficiency test, with correlations ranging from .70 to .78 (see Table 5).

In terms of the achievement tests, two levels of the IEP were included in the analysis for the first research question: Level 2 and Bridge. For each test, a total of 10 points was possible for all speaking tasks; with respect to the Level 2 test, the mean score was highest for the paired task ($M = 8.55$, $SD = 1.54$), followed by the summary task ($M = 7.55$, $SD = 1.23$) and the integrated task ($M = 7.33$, $SD = 1.26$). There was a small correlation between the paired and integrated tasks, with moderate correlations between the summary and paired tasks, as well as between the summary and integrated tasks (see Table 6).

As for the Bridge level test, the mean score was highest for the paired task ($M = 9.42$, $SD = 1.81$), followed by the summary task ($M = 8.39$, $SD = 1.50$) and the integrated task ($M = 8.37$, $SD = 1.67$). There were small correlations between the paired and integrated tasks, as well as between the paired and summary tasks, while there was a high-moderate correlation between the summary and integrated tasks (see Table 7).

The second research question looked at whether practical resources would be different if only two, instead of three, tasks were used in each test. The premise was that if two of these three tasks were overlapping greatly or showing a very weak relationship, then eliminating one would save resources. Time analysis for both the proficiency and

Table 5 Summary of correlations among speaking tasks for IEP proficiency test

Task	1	2	3	*M*	*SD*
1. Narrative	–	.78*	.74*	4.26	1.84
2. Independent		–	.70*	4.05	2.11
3. Integrated			–	3.09	1.95

* $p < .05$

Table 6 Summary of correlations among speaking tasks for IEP achievement test, level 2

Task	1	2	3	*M*	*SD*
1. Paired	-	.47*	−.01	8.55	1.54
2. Summary		–	.44*	7.55	1.23
3. Integrated			–	7.33	1.26

* $p < .05$

achievement tests revealed that the time for designing, administering, and scoring the different speaking tasks varied. As shown in Appendix A, the most time-consuming task (in terms of total hours) in the proficiency test was the integrated task, followed by the narrative task and then the independent task; for both achievement tests, the integrated task was also the most time-consuming task, followed by the paired task and the summary task.

Since one of the goals of this study also was to find out how much money would be saved if a task was eliminated from each test, a cost analysis for the proficiency and achievement tests was deemed necessary. The cost analysis results for the proficiency test (see Appendix B) are based on the testing of around 200 students each time the test is administered, which is five-to-six times per year. Based on the cost analysis for the proficiency test, the integrated task involved the most cost ($1749.00), followed by the independent task ($1465.00) and the picture narrative task ($1430.00). The total cost estimate for all three tasks in the proficiency test was approximately $4644.00.

In terms of the cost of designing, administering, and scoring each task in the achievement test (see Appendix C), the cost per task was less than the cost per task for the proficiency test. The cost estimates are based on testing approximately 30 students once in a given semester. Based on the cost analysis, the integrated task involved the most cost ($451.00), followed by the paired task ($402.00) and the summary task ($309.00). The total cost estimate for all three tasks in each achievement test was approximately $1162.00.

Discussion

Construct representativeness

One aim of the present study was to determine how well each of the tasks represented the aforementioned construct of speaking. The variety of speaking tasks included in the proficiency and achievement tests served to elicit oral speech samples that would provide evidence of students' ability to speak. Those speaking tasks, as they relate to the proficiency and achievement tests, are discussed below.

Table 7 Summary of correlations among speaking tasks for IEP achievement test, bridge

Task	1	2	3	*M*	*SD*
1. Paired	-	.16	−.01	9.42	1.81
2. Summary		–	.68*	8.39	1.50
3. Integrated			–	8.37	1.67

* $p < .05$

Proficiency test

The moderately strong correlation coefficients found among the scores for the three speaking tasks are common for standardized tests, as many large-scale, English language proficiency tests report even higher correlations (ETS, 2011; Zhang, 2008). The tightly controlled test administration conditions and scoring procedures of such tests help to ensure greater reliability of scores. Similar administration conditions and scoring procedures were implemented for our proficiency test, including specialized training for test administrators and raters. In addition, our test included task design specifications and scoring criteria that were modeled after large-scale proficiency tests, such as TOEFL iBT and IELTS. Our efforts to follow the standards and procedures implemented for large-scale tests resulted in acceptable reliability of speaking scores (ranging from .53 to .90).

Based on the findings, each of the tasks contributed something unique to our speaking construct. Because the correlation coefficients were found to be moderately strong (.70 to .78), the task demands for the narrative, independent, and integrated speaking tasks appeared to be related, but adequately independent of each other, thus providing some evidence of convergent validity (see Pae, 2012). This notion is supported by the fact that none of the speaking tasks accounted for more than 60% of the total shared variance. Although the moderately strong correlation coefficients did indicate some degree of overlap between the skills and abilities being assessed by the three tasks, the correlations were not strong enough to suggest that they were assessing exactly the same things. Therefore, we believe that each of the tasks in the proficiency test adequately represented the desired speaking construct.

Achievement tests

For the Level 2 achievement test, the correlation coefficients for the paired/summary tasks and the summary/integrated tasks were found to be moderate and positive (.47 and .44), while the correlation for the paired/integrated task was weak and negative (−.10). In comparison to the larger, positive values found for the proficiency test, the correlations between the three tasks used in the Level 2 test were considerably smaller. This finding is not entirely unexpected, as teacher-made tests typically do not attain the same levels of quality and reliability observed for standardized tests (Burke, 2009; Good, 2008; Zucker, 2003). Therefore, the correlations found among the paired/summary and summary/integrated tasks were considered acceptable for our purposes, while the low correlation for the paired/integrated task was not considered acceptable.

Despite the expectation that correlations among speaking tasks would be lower for the achievement tests, it was not expected that the correlation between the paired/integrated tasks would be much weaker than between the paired/summary and summary/integrated tasks. One potential reason for the low correlation could be that different scoring categories and descriptors were included in the paired-speaking rubric than those included in the rubrics used to score the summary and integrated tasks. Whereas the rubrics for these two tasks had the same three scoring categories (i.e., delivery, content, and language use), albeit with slightly revised descriptors for content, the paired-speaking rubric included scoring categories that were unique to that particular task

(i.e., collaboration, task completion, and style). As a result, raters might have applied the criteria more consistently across the summary and integrated tasks, as the criteria were comparable for both tasks.

Finally, based on the correlation coefficients, the paired-speaking tasks implemented in both achievement tests appeared to tap some unique aspects of the targeted speaking construct. For example, there was a weak (and negative) correlation found between the paired- and integrated speaking tasks for the Level 2 achievement test (−.01). Similarly, for the Bridge achievement test, there was a weak correlation between the paired- and summary speaking tasks (.16) and a weak (and negative) correlation between the paired- and integrated speaking tasks (−.01). In contrast, there were moderate and strong correlations between the summary and integrated tasks for the Level 2 (.44) and Bridge (.68) achievement tests. The low coefficients and minimal shared variance (less than 3% combined) involving the paired-speaking tasks suggest that the speaking abilities required for this task were somewhat unrelated to the abilities required for the summary and integrated speaking tasks.

As Brooks (2009) indicates, paired-speaking tasks are unique from many other speaking tasks, in that they require interlocutors to co-construct their performance, leading to interactional patterns and language variation that are distinctive. This begs the question of whether collaboration and task completion should be assessed in the broader scope of second language proficiency, particularly as part of speaking ability (see Ducasse & Brown, 2009). While it is beyond the scope of this study, future research should consider the centrality of both elements in pair activities and how the operational definitions for collaboration and task completion can be most useful in a scoring scale, and in speaking tests in general.

Practicality

The second aim of this study was to determine the most cost-effective way with which the speaking construct could be adequately represented. As for any testing program, considerations of practicality largely dictate decisions throughout the test development process (Bachman & Palmer, 2010), and given our modest budget for test development, we needed to ensure that an optimal design plan was in place to balance construct coverage and cost. Overall, we found that, given the costs of development for the speaking tasks, decisions about the number of tasks to include in our proficiency and achievement tests had to be made.

For all tests, the integrated speaking tasks required the most time for development, administration, and scoring. As mentioned earlier, the integrated tasks ask students to read/listen to information from two sources and answer a prompt that requires integrating information from the two sources. Locating or creating these sources takes more time than creating a prompt for the independent or paired tasks, and administering the integrated tasks requires more time since students have to listen and read additional input. In addition, training and scoring also take longer time as teachers need to listen to or read the passages involved in the task, familiarize themselves with the key information, and consider different scoring criteria included in the rubric. Despite the budgetary caveats of the integrated tasks, we decided to keep them in our tests. They elicited the types of skills that are integral for success in

an academic setting (Enright et al., 2008), without too much overlap, as demonstrated by their inter-correlations.

Also, it was decided that all of the speaking tasks would be retained in the proficiency test. In looking at the inter-correlations among the speaking tasks, all of them appeared to provide some unique explanation of students' speaking ability. Furthermore, already knowing that the integrated task would be retained, the elimination of either the narrative or the independent task would not save a considerable amount of money over the course of one year: approximately $8580 for the narrative task and $8790 for the independent task. For the amount of money that would be saved by eliminating one task or the other, we felt that keeping the three tasks would provide better coverage of the speaking construct.

As for the achievement tests, the cost of the speaking tasks ranged from $309 to $451, with a total of $1162 per test. The IEP administered 3 tests per semester that were given to 4 different levels, for a total of 12 tests per semester; the estimated cost could be up to approximately $13,944 per year; if this amount is multiplied by the three semesters offered each year at the IEP, the total cost each year would be approximately $41,832. Eliminating one task could save a substantial amount of money annually.

Despite the potential benefits outlined by some research (e.g., Gan, 2012; May, 2009), we decided that the paired-speaking task would be eliminated from the IEP achievement tests. The task correlated poorly with the summary and integrated speaking tasks, suggesting that it was somewhat problematic. This was confirmed in discussions with IEP teachers during several level-specific meetings. Although it was not the focus of the present study, IEP teachers' perceptions regarding the use of the paired-speaking highlighted some issues with grouping and scoring. Several teachers commented that pairing students fairly was difficult, as students' proficiency levels varied within a given class. As a result, students with higher proficiency levels tended to dominate interactions, while students with lower proficiency levels often struggled to contribute meaningful dialogue. Some teachers-raters also indicated that it was difficult to assign scores, particularly in situations such as those described above. On a few occasions, teacher-raters expressed confusion, as well as frustration, when they had to provide the same scores to students working in "unequal" pairs where one student deserved a lower/higher score than the other. Finally, inconsistencies in scoring the paired-speaking task also influenced our decision. Inter-rater reliability estimates were consistently .35 or below, which, even for teacher-made tests, was largely insufficient (Barootchi & Keshavarz, 2002).

Conclusions

Findings of the study indicated that the speaking tasks for the proficiency test appeared to measure the same construct. Similarly, the speaking tasks that were analyzed for the achievement tests seemed to tap into the same construct, with the exception of the paired-speaking task. Accordingly, aside from the paired-speaking task, we believe it is necessary to keep all of the other speaking tasks so that test users can obtain a comprehensive picture of learners' speaking proficiency. Concerning practicality, when tasks are similar in terms of the type of speech that is elicited (e.g., all monologues), as was the case with speaking tasks in the proficiency test, it may be possible to eliminate a

task. Of course, this depends on the purpose of the assessment, the intended *washback* (see Taylor, 2005), and the availability of resources. However, with summative assessments, we would caution test users to include as many tasks as possible to adequately represent the speaking construct. For purposes of summative assessment, it is advisable to include at least two to three tasks so that test users can assess students' speaking ability across different speech act situations, thus helping to inform decisions about students.

On the other hand, if assessment is conducted for formative purposes, then it is possible to include fewer tasks, especially if practicality is an issue, because the objective is likely to give feedback to students based on their task performance. Although we opted to eliminate the paired-speaking tasks for the achievement tests (which serve as summative assessments at our IEP), these tasks are still important for ELLs to participate in, as they promote interactional skills (e.g., turn-taking and negotiation of meaning) that are not typically found in most other speaking tasks, but are nonetheless important in university settings. Therefore, the paired-speaking tasks, while not appropriate for our summative testing purposes, would be quite beneficial for purposes of formative, classroom-based assessment.

Furthermore, it is important to consider the types of tasks that students will be required to perform in the target-language use domain when deciding on tasks to include in a test. For example, stating one's opinion is often required in mainstream North American university classrooms. Therefore, eliminating an independent speaking task should not be a priority. However, while narrating a story may often be necessary in daily life, and to some extent an important skill in academic contexts, it may not always be the highest priority in terms of the kinds of tasks that students are asked to do in university classrooms. Although we kept the narrative task included in the proficiency test, this case study helps to illustrate the importance of prioritizing the needs of an IEP and to implement the assessments accordingly.

Finally, speaking assessments require significant resources to develop, administer, and score. Therefore, it is necessary to consider *a priori* the practicality issues that might arise. Test users/developers need to consider the time, human resources, and money that are available in every phase of test development. Ultimately, how to balance the practicality of an assessment, while also adequately representing the construct of interest, depends on the purpose of the assessment and the decisions that will be made based on the assessment results.

Appendices

Table 8 Summary of time required to design, administer, and score speaking tasks by test type

Test task	Total number of hours
Proficiency	
Narrative	9.00
Independent	4.50
Integrated	13.50
Achievement	
Paired	8.00
Summary	4.50
Integrated	10.00

Table 9 Cost analysis for the proficiency test

Resources	Description	Picture narrative	Independent	Integrated
Personnel	Development (1 test writer)	6 h x $17/h	2 h x $17/h	10 h × 17/h
	Administration (~ 200 students for 10 teachers)	3 mins/student x $17/h	3 mins/student x $17/h	5 mins/student x $17/h
	Scoring (~200 students for 10 teachers)	3 mins/student x $17/h	2 mins/student x $17/h	3 mins/student x $17/h
(subtotal)		$350	$430	$623
Space	Development	100 sq. ft. for 6 h @ .016/h	100 sq. ft. for 2 h @ .016/h	100 sq. ft. for 10 h @ .016/h
	Administration	1500 sq. ft. for 3 h @ .20/h	1500 sq. ft. for 3 h @ .20/h	1500 sq. ft. for 3 h @ .20/h
	Scoring	1200 sq. ft. for 3 h @ .016/h	1200 sq. ft. for 3 h @ .016/h	1200 sq. ft. for 3 h @ .016/h
(subtotal)		$970	$965	$976
Equipment	Development	6 h x $10/h	2 h x $10/h	10 h x $10/h
	Administration (recorders)	30 x $1/unit	30 x $1/unit	30 x $1/unit
	Scoring	2 h x $10/h	2 h x $10/h	2 h x $ 10/h
(subtotal)		$110	$70	$150
Grand total		$1430	$1465	$1749

Table 10 Cost analysis for the achievement test

Resources	Description	Paired	Summary	Integrated
Personnel	Development (2 test writers)	5 h x $17/h	2 h x $17/h	6 h x $17/h
	Administration (~ 30 students for 1 teacher)	3 mins/student x $17/h	3 mins/student x $17/h	5 mins/student x $17/h
	Scoring (~30 students for 2 teachers)	3 mins/student x $17/h	2 mins/student x $17/h	3 mins/student x $17/h
(subtotal)		$162	$94	$196
Space	Development	100 sq. ft. for 6 h @ .016/h	100 sq. ft. for 2 h @ .016/h	100 sq. ft. for 10 h @ .016/h
	Administration	1500 sq. ft. for 3 h @ .016/h	1500 sq. ft. for 3 h @ .016/h	1500 sq. ft. for 3 h @ .016/h
	Scoring	1200 sq. ft. for 3 h @ .016/h	1200 sq. ft. for 3 h @ .016/h	1200 sq. ft. for 3 h @ .016/h
(subtotal)		$140	$135	$145
Equipment	Development	5 h x $10/h	2 h x $10/h	6 h x $10/h
	Administration (recorders)	30 x $1/unit	30 x $1/unit	30 x $1/unit
	Scoring	2 h x $10/h	2 h x $10/h	2 h x $10/h
(subtotal)		$100	$80	$110
Grand Total		$402	$309	$451

Acknowledgments
We would like to thank Dr. Joan Jamieson for her insights and valuable feedback with this project, as well as the PIE at Northern Arizona University for access to its assessment data.

Funding
Not Applicable: No funding was received.

Availability of data and materials
Data will be available upon request. Permission from data owner, the Program in Intensive English, Northern Arizona University, should be obtained first.

Authors' contributions
AB: 50%, SM: 25%, MA: 25%. All authors read and approved the final manuscript.

Authors information (Bios)
Anthony Becker is an Assistant Professor in the English Department at Colorado State University, Colorado, USA. He holds a PhD in Applied Linguistics and his current teaching/research focuses on second language assessment, research methods in applied linguistics, and meta-cognitive aspects of writing. Sawako Matsugu is an Adjunct Lecturer in the Language Center at Rikkyo University, Tokyo, Japan. She Holds a PhD in Applied Linguistics and her research interests are in language testing and assessment with a focus on effects of rater characteristics and scoring methods on speaking assessment. Mansoor Al-Surmi is a Lecturer and a Program Coordinator in the Department of English, Foundation Program, at Qatar University, Doha, Qatar. He holds a PhD in Applied Linguistics and his research interests include investigating theoretical and practical issues in the areas of corpus linguistics and language variation, assessment, and second language acquisition.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Department of English, Colorado State University, 352 Eddy Hall, Fort Collins, USA. [2]Language Center, Rikkyo University, Tokyo, Japan. [3]Department of English, Foundation Program, Qatar University, P.O. Box 2713, Doha, Qatar.

Received: 1 October 2017 Accepted: 20 November 2017
Published online: 02 December 2017

References
Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.
Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.
Barootchi, N., & Keshavarz, M. H. (2002). Assessment of achievement through portfolios and teacher-made tests. *Education Research, 44*, 279–288 https://doi.org/10.1080/00131880210135313.
Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing, 26*, 341–366 https://doi.org/10.1177/0265532209104666.
Burke, K. (2009). *How to assess authentic learning* (5th ed.). Thousand Oaks, CA: Corwin Press.
Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). TOEFL 2000 speaking framework: A working paper. In *TOEFL monograph series, number 20*. Princeton, NJ: Educational Testing Service Retrieved from https://www.ets.org/Media/Research/pdf/RM-00-06.pdf.
Cummins, J. (2008). BICS and CALP: Empirical and theoretical status of the distinction. In B. Street & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., pp. 71–83). New York: Springer Science + Business Media LLC.
Davies, W. M. (2009). Groupwork as a form of assessment: Common problems and recommended solutions. *Higher Education, 58*, 563–584 https://doi.org/10.1007/s10734-009-9216-y.
Douglas, D. (1997). *Testing speaking ability in academic contexts: Theoretical considerations. TOEFL monograph series, number 8*. Princeton, NJ: Educational Testing Service Retrieved from https://www.ets.org/Media/Research/pdf/RM-97-01.pdf.
Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing, 26*, 423–443 https://doi.org/10.1177/0265532209104669.
Educational Testing Service. (2011). *TOEFL iBT research insight: Reliability and comparability of TOEFL iBT scores*. Princeton, NJ. Retrieved from https://www.ets.org/s/toefl/pdf/toefl_ibt_research_s1v3.pdf.
Enright, M. K., Bridgeman, B., Eignor, D., Kantor, R., Mollaun, P., Nissan, S., Powers, D., & Schedl, M. A. (2008). Prototyping new assessment tasks. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 97–143). New York: Routledge Publishing.
French, A. (2003). The change process at the paper level. Paper 5, speaking. In C. Weir & M. Milanovic (Eds.), *Continuity and innovation: Revising the Cambridge proficiency in English examination* (pp. 367–446). Cambridge: UCLES/ Cambridge University Press.
Fulcher, G. (2003). *Testing second language speaking*. London: Pearson Longman.
Gan, Z. (2012). Understanding L2 speaking problems: Implications for ESL curriculum development in a teacher training institution in Hong Kong. *Australian Journal of Teacher Education, 37*, 42–59. http://doi.org/10.14221/ajte.2012v37n1.4.
Good, T. L. (2008). *Twenty-first century education: A reference handbook (volume 2)*. Thousand Oaks, CA: Sage Publications.
Grimm, L. G. (1993). *Statistical applications for the behavioral sciences*. New York, NY: Wiley & Sons.
Hillyard, L., Reppen, R., & Vásquez, C. (2007). Bringing the outside world into an intensive English programme. *ELT Journal, 61*, 126–134 https://doi.org/10.1093/elt/ccm005.
Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.
May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing, 26*, 397–421 https://doi.org/10.1177/0265532209104668.

Nasri, E., & Shokrpour, N. (2012). Comparison of intensive and non-intensive English courses and their effects on the student's performance in an EFL university context. *Eur Sci J, 8*, 127–137 Retrieved from http://www.eujournal.org/index.php/esj/article/viewFile/136/141.
Nazara, S. (2011). Students' perception on EFL speaking skill development. *Journal of English Teaching, 1*, 28–43 Retrieved from https://jetuki.files.wordpress.com/2011/05/3-students-perception-on-efl-speaking-skill-development-pp-28-43.pdf.
North, B. (2000). *The development of a common framework scale of language proficiency*. New York: Peter Lang Publishing.
Pae, H. K. (2012). *Construct validity of the Pearson test of English academic: A multitrait-multimethod approach*. Retrieved from https://pearsonpte.com/wp-content/uploads/2014/07/ResearchNoteConstructvalidityfinal2012-10-02gj.pdf.
Pearlman, M. (2008). Finalizing the test blueprint. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 227–258). New York: Routledge.
Powers, D. (2010). *The case for a comprehensive, four-skills assessment of English language proficiency*. TOEIC compendium study, number 12.2. Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/RD_Connections14.pdf.
Richards, J. C. (2008). *Teaching listening and speaking: From theory to practice*. Cambridge: Cambridge University Press.
Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks* (pp. 167–185). London: Longman.
Stoynoff, S. (2002). Case studies in TESOL practice. *ELT Journal, 58*, 379–393 https://doi.org/10.1093/elt/58.4.379.
Taylor, L. (2005). Key concepts in ELT: Washback and impact. *ELT Journal, 59*, 154–155 https://doi.org/10.1093/eltj/cci030.
Thornbury, S., & Slade, D. (2006). *Conversation: From description to pedagogy*. Cambridge: Cambridge University Press.
Zhang, Y. (2008). *Repeater analyses for TOEFL iBT*. ETS research report (RM-08-05). Princeton, NJ: ETS. Retrieved from https://www.ets.org/Media/Research/pdf/RM-08-05.pdf.
Zucker, S. (2003). *Assessment report: Fundamental of standardized testing*. New York, NY: Pearson Education.