Validity of the YTK Speaking Test:
Construct Validation of a Performance-based English Speaking Test for
Elementary School Students in Japan

KOYAMA Tetsuharu       YUKAWA Emiko

This paper discusses the construct validity of the "YTK Speaking Test," a performance-based language (English) test for students at the elementary school level in Japan. The test was administered to 48 sixth-grade students from a public elementary school in the west region of Japan, and the ratings were analyzed against the convergent, discriminant, and known-groups validity of the test. The results revealed that the YTK speaking test functions as a properly valid language test to measure essential aspects of the students’ communicative competence, especially conversation ability in English, which is expected to be cultivated in elementary school level EFL students in Japan. Some implications and further research questions regarding validity of the test will be discussed.

1. Introduction

Although formal evaluations are not required or encouraged at the elementary school level for English education in Japan, the authors have repeatedly emphasized the importance, necessity, and effectiveness of introducing and utilizing formative evaluations in Gaikokugo-Katsu-sudo (Foreign Language Activities) (Yukawa, Takanashi, & Koyama, 2008; Yukawa, Koyama, & Takanashi, 2009; Koyama, 2009, 2010).

If and when administered, any test must be reliable and valid; a speaking test for students at the elementary school level is no exception. Although it is quite difficult and time consuming to establish reliability and validity for such tests, i.e., speaking or performance tests, introducing into educational environments a test that lacks reliability or validity not only is

---

1 This work is partially based on the paper presented at the ninth conference of the JES (Japan Association of English Teaching in Elementary School) in July, 2009, at Tokyo Gakugei University (Koyama, Yukawa, & Takanashi, 2009). Also, this work is a part of a larger project (i.e., the YTK project) which has been supported by MEXT: Grant-in-Aid for Scientific Research (B) (20320087).

2 Professor at Ritsumeikan University. Also a part-time lecturer at the graduate program in Applied English, Kyoto Notre Dame University. The primary investigator of the grant project (YTK Project).
useless, but could be rather harmful. If the results, especially numerical scores, are reported without adequate interpretations, an unfortunate spiral could be a consequence: in the worst scenario, a curriculum could be re-designed in a way that encourages the students to reach a certain level (i.e., score) on the test even without full and appropriate understanding of what those scores possibly indicate.

This may sound too obvious to state. Very unfortunately, however, such a negative spiral does not seem rare in educational environments. Thus, researchers in the field have repeatedly emphasized the importance of establishing “usefulness” of language tests. For instance, Bachman and Palmer (1996) contended:

“The most important consideration in designing and developing a language test that is the use for which it is intended, so that the most important quality of a test is its usefulness. ..... By stating the obvious and by questioning it, we wish to point out that although usefulness is of unquestioned importance, it has not been defined precisely enough to provide a basis for either designing and developing a test or for determining its usefulness after it has been developed” (p.17).

According to the authors, usefulness of a language test can be a function of six important qualities: reliability, construct validity, authenticity, interactivenss, impact, and practicality (ibid, p.18) (see figure 1). These six qualities are without doubt equally important, and a language test must be evaluated in terms of their combined effect on the overall usefulness. The authors of this paper agree with Bachman and Palmer (1996) that two of the qualities - reliability and validity - are most important because “these are the qualities that provide the major justification for using test scores - numbers - as a basis for making inferences or decisions” (p.19), and therefore find it necessary to start with checking the validity of the YTK speaking test before discussing overall usefulness of the test. To that end, a series of validity investigations of the test have been conducted. In the sections that follow, this paper reports the results of the construct validation and then discusses whether the YTK speaking test can function as a “useful” test of communicative competence in English among students at the elementary school level in Japan.
2. Background and Research Hypotheses

2.1. YTK Speaking Test

The YTK speaking test was designed and developed to assess the essential aspects of communicative competence that are expected to be developed in elementary school students (Yukawa, Takanashi, & Koyama, 2008, 2009). As of the end of the academic year of 2010, a total of 734 students had participated in the test, and the results and implications were discussed elsewhere. Below, the purposes and design of the test is briefly described (for the comprehensive review of the test, see Yukawa, Tanakashi, & Koyama, 2009).

Communicative competence

Definitions and components of communicative competence have been discussed extensively in the literature of applied linguistics (Bachman and Palmer, 1996; Hymes, 1972; Swain and Canale, 1980; among others) as well as in the field of interpersonal communication studies (Spitzberg, 1995, 2003; Spitzberg & Cupach, 1989, 2002; among others), and all seem to have agreed that communicative competence should not be seen to base solely on “linguistic competence” (i.e., phonological and syntactical knowledge), but should also include other competences crucially linked to language use, namely competences often referred to as ‘illocutional,’ ‘socio-linguistic,’ or ‘strategic’ competences (Bachman and Palmer, 1996; Hymes, 1972; Swain and Canale, 1980). Based on this view, the YTK speaking test was developed to measure both the linguistic aspects (i.e., phonology and vocabulary & syntax) and the interactive aspects (i.e., attentiveness, expressiveness, and conversational management skill) of communicative competence among elementary school students.

In order to measure such ‘interactional’ aspects of communicative competence, the test was designed to observe performance, not just static knowledge of language skills. The oral test/task format, especially paired/group performance testing, has been recommended as an assessment that is more valid and learning-oriented in different frameworks: Teacher-Based Assessment (Davison and Leung, 2009), Dynamic Assessment (DA, Poehner, 2009), and others.
(Cameron, 2001; Luoma, 2004). In these frameworks, communication is seen as a product of co-construction by all the participants, and thus communication abilities are not regarded residing within individuals (MacNamara, 1997). The YTK speaking test is generally based on these perspectives.

**Scaffolding**

The test was also designed to allow “scaffolding” (Cameron, 2001). Considering the low English level of students English in elementary schools especially in Japan, scaffolding is necessary for a test to reveal not only what students can do without help at that moment, but to gain an assessment of what they will be able to do in the future (Vygotsky, 1998). Thus, participants are paired so that they may help each other (and/or ask the tester for help) when necessary. Furthermore, in contrast to one-on-one interviews, the students’ stress level can be lowered in the paired interviews (Luoma, 2004).

**Scoring procedures**

The performance was videotaped and rated afterwards by trained raters in five areas: (1) pronunciation, (2) vocabulary and syntax, (3) attentiveness, (4) expressiveness, and (5) conversational management, all on a rubric 4-point scale (4 = highest, 1 = lowest) (for the detailed coding criteria, see Yukawa, Koyama, & Takanashi, 2010).  

2.2. Validity Concerns and Research Assumptions

Construct validity concerns the theoretical relationship of one variable to another (Cronbach & Meehl, 1955), and is defined as “the extent to which a measure behaves the way that the construct it purports to measure should behave with regard to the established measures of other constructs” (DeVill, 1991, p.46). In social science in general, construct validity is categorized into four different types: (1) intercorrelations validity, (2) convergent validity, (3) discriminant validity, and (4) known-groups validity (Brown, 1996; Singleton & Straits, 2004). Intercorrelations validity concerns correlations between the target variable and other theoretically related variables: they should be strongly correlated. Convergent validity requires consistency across indicators and different methods of measurement: different measures of the same concept should be strongly correlated (i.e., multimethod validity). Discriminant validity concerns correlations among theoretically unrelated variables: they should not be strongly correlated (i.e., multitrait validity). Known-groups validity refers to the consistent differences

---

3 The score was not shown to the students; they were told that accomplishing the task itself is their goal.
among known groups: when one group is expected to be different from another group in terms of a construct, the groups’ responses should be different.

Amongst the four validity types, this study investigated the three types of validity of the YTK speaking test: convergent, discriminant, and known-groups validity. The reason for this was that theoretically interrelated constructs to communicative competence (i.e., the YTK speaking test scores) were not available at that time. Considering the available resources (i.e., data from tests/measurements administered), the authors built several assumptions (hereafter, called hypotheses) regarding three types of validity of the YTK speaking test:

**Convergent validity**

H1: The YTK Test ratings should positively correlate with the scores from different measures of the same aspect of communicative competence.

**Discriminant validity**

H2: The YTK Test ratings should reflect unique and specific aspects of communicative competence (i.e., speaking/conversational competence), but not others (e.g., listening ability)

**Known-groups validity**

H3: The YTK Test ratings should properly reflect expected changes (i.e., improvement) in communicative competence levels across time.

H4: The YTK Test ratings should reflect the difference in communicative competence levels among different year groups.

3. Methods

3.1. Construct Validations

As outlined in Campbell and Fiske (1967), a multitrait-multimethod (MTMM) convergent-divergent design would be one of the ideal frameworks to test against the overall construct validity of a language (or any kind of) measurement (Brown, 1996). Clifford (1981) and Bachman & Palmer (1981), for instance, discussed the effectiveness of the MTMM design for construct validation of several language tests, concluding that the Campbell-Fiske scheme should be incorporated especially in data collection designs.

In Bachman & Palmer (1981), confirmatory-factor-analytic procedures were recommended for formulating and testing hypotheses regarding the target traits.
For practical reasons (e.g., limited access to the students, etc.), however, it was impossible to build a full-scale Campbell-Fiske MTMM design for the current study. Instead, convergent and discriminant validations were conducted independently to examine each research assumptions regarding construct validity of the YTK speaking test.

Convergent Validity (H1)

To test the first hypothesis (convergent validity), a correlational analysis was conducted. The YTK speaking test scores were correlated with the scores from two other measures: the interviewers’ ratings on the students’ communicative competence, and the instructors’ ratings of the students’ communicative competence (i.e., multimethod validation).

Discriminant Validity (H2)

To test the second hypothesis, a correlational analysis was conducted. The YTK speaking test scores were correlated with the scores from another measure (the YTK Listening Test) that assessed receptive skills of the same students (i.e., multitrait validation).

Known-groups Validity (H3 & H4)

To test the third and forth hypotheses, a repeated-measure t-test was conducted between the YTK speaking test scores from two different groups. For the third hypothesis, the scores from the students’ first and second sessions were compared. For the fourth hypothesis, the scores of the target students (2009 group) were compared against the average scores from the previous year (2008 group) (i.e., different-groups validation (Brown, 1996), or known-groups validation (DeVellis, 1991)).

3.2. Administration of the YTK Speaking Test

Participants and Background

A total of 48 sixth-grade students (two classes) from a public elementary school in the west region of Japan participated in the YTK speaking test, in March, 2009. The same group of students also participated in the YTK listening test in the same year5.

Preparation

As mentioned earlier, the YTK speaking test was introduced to the students not as a test, but rather as the final project/activity to wrap up the course. It was introduced under the ti-

5 Both tests were not administered particularly for this study; rather, they were introduced as parts of the formative evaluation in the course.
tle “Let’s Talk,” providing the students with a rare and precious opportunity to have real and hopefully exciting experience of interacting with native speakers of English.

Before participating in Let’s Talk, the students had a total of five “preparation” sessions during the five weeks period. In each session, the students were divided into small groups consisting of five to six members each, and practiced asking and answering several questions they would use/hear in the upcoming conversation session. Throughout the preparation sessions, each group was attended by one or more instructors (i.e., trained graduate students, supervised by one of the authors); the sessions were conducted in an interactional style, rather than a “drill” or “try-to-memorize-the-assigned-expressions” style.

**Procedure and Design**

The students were paired up and told to walk into a room where an English native speaker (a total stranger) was waiting. After a greeting, the paired students carried conversation with the interviewer, using English phrases they knew. The students were encouraged to maintain the conversation in English as much as possible, but were also told that it was totally acceptable to use nonverbal or even Japanese expressions if and when necessary. Furthermore, the students were allowed to help each other, and the interviewer was also instructed by the researchers to help the students as they would normally do in natural conversations (thus, “scaffolding”). Each conversation was timed, and the assistant signaled when it reached 3 minutes. On hearing/seeing the signal, the interviewer naturally ended the conversation, said goodbye, and had the students leave the room. All the students had two conversation sessions on the same day, with different interviewers and different partners. Before and between the conversation sessions, the students remained in a waiting room (regular classroom) with fellow students. No specific directions were given while they were waiting; the students were free to talk about or prepare for their interviews.

**3.3. Measurements**

Four different measures were used to investigate into the construct validity of the YTK Speaking Test: (1) the YTK Speaking Test scores, rated by the trained coders, (2) the interviewers’ assessments of the students’ ability in conversing in English, (3) the instructors’ (i.e., the graduate students who attended the preparation sessions) assessments of the

Some students took the role of “interviewers” and actively engaged in the activity. Again, the sessions were not meant to be “preparation for the test,” but rather were conducted as necessary steps toward their final project in which they the students were expected to “welcome” the guests and enjoy conversing with them.
students’ attitudes and competence, and (4) the YTK Listening Test scores.

(1) YTK speaking test scores

All the conversation sessions were video-recorded and transformed into movie files visible on a personal computer. Three coders (i.e., the authors and the assistant, all trained) rated the conversations on the five scales described earlier (see Appendix A; for the detailed rating procedures and inter-coder reliability issues, see Yukawa, Takanashi, and Koyama, 2009). Each coder rated all the conversations, and Cohen’s Kappa (inter-coder reliability index) was calculated for each scale: .56 for the Phonology scale (66.67% exact matching among the three), .51 for the Syntax (63.12% exact matching), .70 for the Attentiveness scale (77.66% exact matching), .44 for the Expressiveness scale (57.80% exact matching), and .55 for the Conversational Management scale (65.96% exact matching). When the two ratings matched but the other rating deviated from the other two by one point, the matched score was accepted; when there was two point difference anywhere among the ratings, the coders met and discussed until all agreed to accept one rating.

Between the two conversation sessions, the higher score was adopted for the main analyses, because the YTK test was designed to measure the maximum ability possible of elementary school children.

(2) Interviewers’ assessments

The interviewers rated the conversation on two 4-point scales: conversational management ability and overall impression (see Figure 2). To grasp very intuitive and “natural” impressions of the conversation by the participant (i.e., the interview), the interviewer was asked to rate the session immediately after the students left the room.

1. Conversational Management
The student understood her/his conversational role (as a speaker or a hearer), and acted accordingly to maintain conversational flow. (keys: nodding, turn-taking, managing topics to avoid awful silence, etc.)

1  2  3  4

2. Overall Impression
The student was a pleasant conversation partner and I could enjoy interacting with her/him.

1  2  3  4

Figure 2. Scales used by the interviewers
(3) **Instructors’ assessment**

The instructors who taught the “preparation session” were asked to rate the students overall attitude toward the preparation sessions and their communicative competence, both on a 4-point scale (see Figure 3).

1. 学習意欲・態度
   （意欲をもって、学ぶ態度で自分なりに精一杯とりくんでいたかどうか）
   
   ![Figure 3](image)

2. 話す力
   （会話の際に相手がいうことを聞き取り、自分がちゃんとしゃべるだけの英語力を持っているかどうか：特に語彙、構文、発音等）

   ![Figure 3](image)

(4) **YTK listening test scores**

The same group of students participated in the YTK listening test in 2009, which was introduced to the students as “Eigo-chikara-dameshi (Try Out Your English),” but not as a test in a regular sense. The YTK listening test was designed and developed to measure the receptive aspects of communicative competence among elementary school children, including Bachman and Palmer’s (1988) model of “textual competence” and “grammatical competence” (Yukawa, Takanashi, Koyama, 2008; among others). This test comprised of 37 questions, which tested whether the students can comprehend the vocabulary of such topics as the days of the week, months of the year, telling time, weather, numbers, countries, colors, sports, food, animals, school subjects, directions, shapes, and occupations in English. Furthermore, it tests whether the students can answer when asked directions, make an order at a fast food restaurant, or understand classroom commands during lesson time. It also tests whether the students can recognize English sounds and letters, and whether they can understand easy children’s storybook (16 pages of one short sentence of 4 to 7 words per one page). The test is recorded on a DVD with both still pictures and animations. Scores were obtained on a 80-point scale.

4. **Results & Discussion**

4.1. **Descriptive Statistics of Four Measurements**

On the YTK Speaking Test, the students scored on average 2.88 (out of 4, SD = .64), with average scores ranging .50 across the components (on Phonology, M = 2.84, SD = .58; Vocab-
ulary and Syntax, $M = 2.64, SD = .69$; Attentiveness, $M = 3.14, SD = .67$; Conversational Management, $M = 2.90, SD = .84$). The interviewers’ ratings of the students’ communicative competence during the interviews were on average 3.08 ($SD = .73$) on the Conversational Management scale, and 3.55 ($SD = .50$) on the Overall Impression scale. The instructors’ (i.e., graduate students who functioned as group leaders in preparation sessions) ratings were on average 3.70 ($SD = .62$) on the attitude scale and 3.81 ($SD = .45$) on the competence scale. Although ceiling effects were observed with the instructors’ ratings on both scales, as well as the interviewers’ rating on the overall impression scale, the authors decided to include all the scales for the analyses considering the explorative nature of this study. The students scored on the YTK Listening test on average 68.35 ($SD = 8.43$), which is equivalent to 79.81%. See Table 1 for the details.

<table>
<thead>
<tr>
<th></th>
<th>$N$</th>
<th>$M$</th>
<th>$SD$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>YTK Speaking Test</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coder Ratings</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phonology</td>
<td>48</td>
<td>2.90</td>
<td>.43</td>
</tr>
<tr>
<td>Vocab &amp; Syntax</td>
<td>48</td>
<td>2.69</td>
<td>.59</td>
</tr>
<tr>
<td>Attentiveness</td>
<td>48</td>
<td>3.21</td>
<td>.50</td>
</tr>
<tr>
<td>Expressiveness</td>
<td>48</td>
<td>2.96</td>
<td>.74</td>
</tr>
<tr>
<td>Conv. Management</td>
<td>48</td>
<td>2.98</td>
<td>.73</td>
</tr>
<tr>
<td>Average</td>
<td>48</td>
<td>2.95</td>
<td>.49</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$N$</th>
<th>$M$</th>
<th>$SD$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interviewers’ Ratings</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conv. Management</td>
<td>48</td>
<td>3.08</td>
<td>.73</td>
</tr>
<tr>
<td>Overall Impression</td>
<td>48</td>
<td>3.55</td>
<td>.50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$N$</th>
<th>$M$</th>
<th>$SD$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instructors’ Ratings</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attitude</td>
<td>47</td>
<td>3.70</td>
<td>.62</td>
</tr>
<tr>
<td>Ability to Converse</td>
<td>47</td>
<td>3.81</td>
<td>.45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$N$</th>
<th>$M$</th>
<th>$SD$</th>
</tr>
</thead>
<tbody>
<tr>
<td>YTK Listening Test</td>
<td>46</td>
<td>63.85</td>
<td>8.43</td>
</tr>
</tbody>
</table>

Notes: YTK Speaking Test measures and Instructors’ ratings were all on 4-point scales; TYK Listening Test was on 80-point scale.

4.2. Convergent validation (Hypothesis 1)

To establish convergent validity, the YTK speaking test ratings should positively correlate with the scores from a different measure of the same aspects of communicative competence: the interviewers’ ratings on the students’ communicative competence, and the instructors’ ratings of the students’ communicative competence.

The hypothesis was partially supported. The correlational analysis (Pearson correlation coefficients) revealed strong positive and statistically significant correlations between the average YTK speaking test score and the Interviewers’ ratings on communicative competence: $r$
Validity of the YTK Speaking Test: Construct Validation of a Performance-based English Speaking Test for Elementary School Students in Japan

= .68 (p < .001) with the Conversational Management rating, and r = .56 (p < .001) with the Overall Impression rating (see Table 2 for the details). However, only low to medium correlations were found between the TYK speaking test scores and the Instructors’ ratings: r = .25 (p < .05, one-tailed) with the Attitude rating, and r = .32 (p < .05) with the Competence rating (also see Table 2 for the details).

Considering that the interviewers were not trained to professionally rate students’ performance in English, their ratings are seen to reflect intuitive and sincere impressions that any real communicator would personally make during and after conversation with the students. In terms of measurement methods, therefore, the interviewers’ ratings are apparently different from the YTK speaking test scores, which were obtained through scientific coding procedures by trained coders. The strong correlations between the two ratings survived a multimethod validation, allowing us to claim that the YTK speaking test, and the Interviewers’ ratings, can validly measure elementary school students’ communicative performance.

The reason for the weak correlations of the Phonology rating with other two ratings is unknown. It is only speculated that either the interviewers were not paying much attention upon the phonological aspect of the students’ performance, or the degree of phonological ability had no impact on the interviewers’ perception of the students’ communicative competence. It is also worth noting that the correlations between the Vocabulary & Syntax rating and the Interviewers’ ratings were much smaller than the correlations for the other YTK rating scales: r = .45 (p < .01) for the Management rating; r = .30 (p < .05) for the overall impression rating, while other correlation coefficients ranged from .54 to .74 (all p < .001). This by no means suggests that the Phonology or Vocabulary/Syntax aspect of communicative

7 All the significance testes were two-tailed, unless stated otherwise as in this case.

Table 2. Multimethod convergent validation

<table>
<thead>
<tr>
<th>YTK Speaking Test Coder Ratings</th>
<th>Instructors’ Ratings</th>
<th>Interviewers’ Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Attitude</td>
<td>Competence</td>
</tr>
<tr>
<td>Phonology</td>
<td>.07</td>
<td>.27</td>
</tr>
<tr>
<td>Vocab &amp; Syntax</td>
<td>.23</td>
<td>.27</td>
</tr>
<tr>
<td>Attentiveness</td>
<td>.24</td>
<td>.32</td>
</tr>
<tr>
<td>Expressiveness</td>
<td>.32*</td>
<td>.25</td>
</tr>
<tr>
<td>Conv. Management</td>
<td>.12</td>
<td>.23</td>
</tr>
<tr>
<td>Average</td>
<td>.25</td>
<td>.32</td>
</tr>
</tbody>
</table>

Notes: N = 36; + p < .05 (one-tailed); * p < .05; ** p < .01; *** p < .001
competence is less important, but at least hints that other aspects often referred to as strategic competence (Bachman & Palmer, 1996; Swain & Canale, xxxx; among others) or interpersonal communication skills (Spitzberg, 1995, 2003; Spitzberg & Cupach, 1989, 2002) contribute more in making the conversation pleasant and smooth. Further research is required.

The instructors’ ratings only weakly correlated with the YTK speaking test scores presumably for two reasons. For one, their ratings were based on the sum of all the students’ performance during the five-week preparation period, which may or may not reflect the same construct the YTK speaking test measure (i.e., the students’ communicative competence). They might have reflected, for instance, attitudes toward the course or the instructor(s), strategies or skills in approaching class tasks in general, etc., in addition to communicative competence. For the other, the instructors rated the students’ performance based on their recollections of the interactions they had with the students approximately two weeks prior. Although the degree to which the time lag affected their recollection is unknown, it is easily speculated that their ratings were not as accurate as the interviewers’.

4.3. Discriminant validation (Hypotheses 2)

**Hypothesis 2: Multitrait validation**

It was assumed that a correlational analysis would yield only low to medium correlation coefficients, as productive/interactional skills and receptive/listening skills are theorized to be somewhat distinctive, although related. To test the assumption, a multitrait validation was conducted: the YTK speaking test scores were correlated with the scores from another measure (the YTK Listening Test) that assessed receptive skills of the same students.

The results confirmed the assumption: the overall correlational coefficient was .30 ($p < .05$) and others ranged from .17 to .30 (see Table 3). It is therefore plausible to claim the YTK speaking test establishes discriminant validity as a measure of speaking and conversational competence among elementary school students.

<table>
<thead>
<tr>
<th>YTK Speaking Test Coder Ratings</th>
<th>YTK Listening Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonology</td>
<td>.27*</td>
</tr>
<tr>
<td>Vocab &amp; Syntax</td>
<td>.28*</td>
</tr>
<tr>
<td>Attentiveness</td>
<td>.17</td>
</tr>
<tr>
<td>Expressiveness</td>
<td>.24</td>
</tr>
<tr>
<td>Conv. Management</td>
<td>.23</td>
</tr>
<tr>
<td>Average</td>
<td>.30*</td>
</tr>
</tbody>
</table>

Notes: $N = 46$; $+ p < .05$ (one-tailed); $^* p < .05$
Also, the scores on Attentiveness, Expressiveness, and Conversational Management, which are seen as unique and important aspects of interactional competence, were not correlated with the YTK listening score (correlational coefficients were .17, .24, and .23, respectively, all n.s.). This result further validates the YTK speaking test in its power of measuring interactive skills, not merely static knowledge, of the students. The YTK speaking test and the YTK listening test can and should be used independently to validly measure different aspects of communicative competence among elementary school students.

4.4. Known-groups validation (Hypotheses 3 & 4)

Hypothesis 3: The same group across-time validation

The students were assumed to perform better in their second conversation session, because the students had had virtually no opportunity to talk with native speakers of English other than their instructor (ALT) prior to the project. The results from a comparison of the scores from two groups (i.e., repeated-measure t-test) supported this hypothesis. The students scored higher in their second trials: overall, $M = 2.73$, $SD = .48$, for their first conversation session, and then $M = 2.88$, $SD = .47$, for the second session, $t_{(42)} = 2.39$, $p < .05$ (see Table 4 for the details).

Table 4. The same group across-time validation (repeated measure t-test)

<table>
<thead>
<tr>
<th>YTK Speaking Test Coder Ratings</th>
<th>First session</th>
<th>Second session</th>
<th>t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$</td>
<td>$SD$</td>
<td>$M$</td>
</tr>
<tr>
<td>Phonology</td>
<td>2.73</td>
<td>.49</td>
<td>2.88</td>
</tr>
<tr>
<td>Vocab &amp; Syntax</td>
<td>2.45</td>
<td>.58</td>
<td>2.63</td>
</tr>
<tr>
<td>Attentiveness</td>
<td>3.08</td>
<td>.45</td>
<td>3.14</td>
</tr>
<tr>
<td>Expressiveness</td>
<td>2.69</td>
<td>.74</td>
<td>2.84</td>
</tr>
<tr>
<td>Conv. Management</td>
<td>2.67</td>
<td>.77</td>
<td>2.93</td>
</tr>
<tr>
<td>Average</td>
<td>2.73</td>
<td>.48</td>
<td>2.88</td>
</tr>
</tbody>
</table>

Notes: $N = 43$ (pairs); $+ p < .05$ (one-tailed); $* p < .05$

Hypothesis 4: Different-groups validation

In the target school, the YTK speaking test (“Let’s Talk”) was first administered in March, 2008, at the end of the academic year of 2007. The students’ performance levels in March, 2008 and those in March, 2009 were expected to be somewhat different for several reasons. First, feedbacks based on the first YTK speaking test results were given to the instructors, and the instructors attempted to revise their new curriculum or improve their teaching methods in the following year. Second, and most importantly, the contents of the preparation sessions differed significantly between the year 2007 (the first year) and the year
2008 (i.e., second year). In the first year (March, 2008), the students in the preparation lessons were only expected to speak about themselves without anyone interviewing them. In the second year (March, 2009), the students prepared for ‘a group introduction presentation,’ and within this task-based teaching unit, the students were encouraged to even interview each other and initiate a conversation asking questions rather than just speaking about themselves as monologues (for the detailed review of the preparation sessions, see Yukawa, Koyama, & Takanashi, 2010). Although it is difficult to predict the exact impact of those educational changes, the students in the 2009 group were expected to perform better than those from the 2008 group, because in 2009 the students had learned how to contribute to the conversation by taking some initiative rather than just waiting for the interviewer ask them questions.

The statistical analysis (t-test) yielded affirmative results: the students from the 2009 group scored on average 2.95 (SD = .49) while those from the 2008 group scored 2.47 (SD = .49) on average (t(94) = 4.76, p < .001).

<table>
<thead>
<tr>
<th>YTK Speaking Test Coder Ratings</th>
<th>2008 M</th>
<th>2008 SD</th>
<th>2009 M</th>
<th>2009 SD</th>
<th>t(94)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonology</td>
<td>2.79</td>
<td>.58</td>
<td>2.90</td>
<td>.43</td>
<td>1.00</td>
</tr>
<tr>
<td>Vocab &amp; Syntax</td>
<td>2.29</td>
<td>.58</td>
<td>2.69</td>
<td>.59</td>
<td>3.31**</td>
</tr>
<tr>
<td>Attentiveness</td>
<td>2.73</td>
<td>.68</td>
<td>3.21</td>
<td>.50</td>
<td>3.94***</td>
</tr>
<tr>
<td>Expressiveness</td>
<td>2.31</td>
<td>.62</td>
<td>2.96</td>
<td>.74</td>
<td>4.61***</td>
</tr>
<tr>
<td>Conv. Management</td>
<td>2.23</td>
<td>.69</td>
<td>2.98</td>
<td>.73</td>
<td>5.12***</td>
</tr>
<tr>
<td>Average</td>
<td>2.47</td>
<td>.49</td>
<td>2.95</td>
<td>.49</td>
<td>4.76***</td>
</tr>
</tbody>
</table>

Notes: N = 48; * p < .05; ** p < .01; *** p < .001

4.5. Other validity consideration: consequential validity (Washback)

Another validity consideration is briefly discussed in this section, although not thoroughly, in order to provide further supports for the overall validity of the test. Consequential validity of a test concerns positive or desired effect of the test implementation. Performance tests, if properly developed, are expected to be consequentially valid; because performance-based tests can provide great closeness to real-world tasks, "class time spent on preparation of students for performance on the assessment tasks is thus preparation for the real-world tasks which they are simulating” (McNamara, 1996, p.22). Also, performance-based tests can have beneficial impact on the teaching: such an impact is labeled as washback effect and is seen

---

8 Washback effect can be negative, of course, if the test is not properly designed and developed (i.e., if not reliable or valid).
as an important aspect of the consequential validity of language tests (Messick, 1989; Wesche, 1987).

This has been the case with the YTK speaking test, although it may be premature to claim so in scientific terms. As already described in the previous section, the curriculum and teaching improved significantly in the following year presumably at least partially due to the administration of the YTK speaking test. Although only through subjective observations, the authors of this paper witnessed positive surprise and great satisfaction in the instructors who observed their students’ performance in “Let’s Talk,” which seemed to motivate and encourage them to improve their curriculum/teaching in the following year. Furthermore, it was more than apparent that the majority of the students enjoyed and felt great pleasure in their somewhat successful communication with native speakers of English, which the authors believe has a very positive influence on the students’ motivations and willingness to study English in their junior high school courses (for further and "scientific" discussion on the predictive validity of the test, see Koyama, Yukawa, & Takanashi, 2011; Yukawa, Koyama, & Sugimoto, 2010; Yukawa, Koyama, & Takanashi, 2010).

5. Conclusion

This study has clearly shown that the YTK speaking test functions as a valid performance test for elementary school level learners of English. A series of construct validations (i.e., multimethod validation, multitrait validation, same group across-time validation, and different groups validation) confirmed that the test establishes the convergent, the discriminant, and the known groups validity necessary for a test to be useful.

With respect to the limitations of the study, the authors recognize that the sample size was small and non-representative, and therefore it is still hasty to conclude that the test is valid for the entire population. Also, as mentioned earlier, this study only employed several independent convergent and discriminant validation of the test. It is highly suggested that a future study should design a full-scale MTMM validation, i.e., fully crossed method-by-measure matrix (DeVellis, 1991), to investigate the construct validity of the test.

There are several implications of the study. With its convergent validity established, the authors recommend the use of the YTK speaking test or a similar performance test to measure elementary students’ communicative competence in English. When the purpose of Gai-koku-go-katsudo-no-ijken (Foreign language activity) is to “cultivate foundations of communicative competence” in English (MEXT, 2009), it is imperative to see if and how such competence can be and is cultivated in the classes, and the YTK speaking test can be a very “use-
ful” measure. Also, with the discriminant validity established, the YTK speaking test should be administered along with the YTK listening test. Combined, these tests can validly grasp the wide range of components of communicative competence.

References
Koyama, T. (2010). Shoogakko eigo de sodate eru “communication nooryoku no soji.” (Foundations of communicative competence in English education at the elementary school level). In E. Yukawa, & Y. G., Butler (Eds.), Shoogakko eigokatsudo hissyuka no tame no curriculum, kyozai, oshiekata (Curriculum, teaching materials, and teaching methods for the
newly required English Activities at the elementary school level), pp. 17-41. MEXT Grant-in-aid-for-scientific-research (B) Report.


*Language Testing, 4*(1), 28-47.

