# Reevaluating the Test Specifications of an Oral Proficiency Test

| | |
|---|---|
| 著者名(英) | Gene Thompson |
| journal or publication title | The journal of Kanda University of International Studies |
| volume | 21 |
| page range | 233-260 |
| year | 2009-03-31 |
| URL | http://id.nii.ac.jp/1092/00001267/ |

# REEVALUATING THE TEST SPECIFICATIONS OF AN ORAL PROFICIENCY TEST

Gene Thompson

## Abstract

In any language testing situation, the reevaluation of the test construct and specifications should be an ongoing exercise which parallels the successes and setbacks of validation process. This paper outlines and discusses an attempt to reevaluate the test specifications, and particularly redefine the construct, of an oral subtest of an in-house English language proficiency test from a University in Japan. The paper begins with background to the test and concepts of speaking, followed by an attempt to clarify aspects of the test task and construct by relating them to models and frameworks of language ability and use from the literature. While a draft set of specifications is detailed, a data-based approach to construct definition and rating scale design is suggested as a complement to the theory based approach followed in this paper, in order to iteratively consider the reevaluation of the test construct from both theoretical and empirical approaches.

## 1.0 Introduction

Good tests have clearly defined specifications; construct, assessment, and task (Luoma, 2004). Taken together these specifications constitute the blueprint for the test, and detail the rationale for the test construct with how it is operationalized through the task and rating procedures employed. Furthermore, test design is an evolving process that requires the revisiting and revision of test specifications

(Chauhoub-Deville & Turner, 2000; Fulcher, 1996; Luoma, 2004).

The purpose of this paper is to consider and critically discuss some of the fundamental issues of principled test design and relate them to the test specifications, and specifically construct definition, of the speaking subsection of an existing in-house proficiency test from a University in Japan; the Kanda English Proficiency Test (KEPT). Since its inception in 1989, very little detailed reevaluation of the test specifications of the oral component has been carried out in relation to developments in speaking or speaking assessment in the field of applied linguistics, and what developments and changes that have been made are not documented well with yearly 'test reports' only being published from 1997. Furthermore, there is little remaining evidence of the initial test design process with no 'original Specs' that current test designers can add to, revise, or otherwise revisit. Sadly, there has been no maintenance of a detailed catalogue of the test design process or of periodical changes to the test construct, rating criteria, or task design. The process of maintaining this catalogue is referred to by Luoma (2004:118) as a 'history file', which can be an important addition to the ongoing test development and validation process as it "encourages self monitoring and qualitative development" and provides a useful resource for sharing and maintaining the collective work on a test over time. For the KEPT, as task and assessment specifications of the test have been adjusted, and as different models of speaking ability have been developed, detailed consideration or reinterpretation of the test construct has lagged. Therefore, the specific purpose of this paper is to attempt to identify and critically discuss which elements of oral proficiency, as it is currently understood, are essential or most useful for helping the KEPT oral to maximize the extent to which we can make the kinds of interpretations about students' oral proficiency that we would like to make about our students in our context.

## 2.0 Towards definitional clarity of speaking

### *2.1 A different view of the speaking skill: Orally assessing proficiency or assessing speaking?*

Bachman and Palmer (1996: 75) criticize 'skills' based distinctions as they fail to account for the idea that "language use takes place, or is realized, in the performance of specific language use tasks", and that therefore

> rather than attempting to define 'speaking' as an abstract skill… it is more useful to identify a specific language use task that involves the activity of speaking and describe it in terms of its task characteristics and the areas of language ability it engages (p.76).

The implication of this 'interactionalist' approach to construct definition, where the definition of the language 'skill' being elicited depends on the interaction between the characteristics of the task and the language ability of the candidates (see Messick 1989, for more) is that the different language use situations employed in the assessment determine the area of language ability required. The point here is that underlying language proficiency is seen as relatively constant and that for second language users; 'strategic competence' mediates individuals' linguistic knowledge with the language use situation, recognizing the importance of context not only in the teaching and using of language, but also the assessment of language. Therefore, given this idea, a speaking test such as the KEPT oral might be better viewed as an oral proficiency test in x context, rather than a general 'speaking' test.

### *2.2 Conversation*

Kormos (1999: 165), in an examination of the relative benefits of using role-play as a test method versus interview, provides an excellent working definition of what constitutes a conversation relevant with respect to the KEPT oral task. Drawing together the ideas of Jones and Gerard (1967), Silverman (1973) Sacks et al. (1974), Goffman, (1976), Kress and Fowler (1979), and Van Lier (1989), Kormos defines conversation as:

> "an unplanned face-to-face interaction with unpredictable sequence and outcome in which the rights and duties of the interactants are equally distributed and in which speakers turns are reactively or mutually contingent"

Simply put, participants don't know where a conversation will go, but nominally (at least) everyone involved has the opportunity to influence the outcome and direction of the interaction. Furthermore they are all responsible for reacting to the utterances of the other participants (e.g. turn taking, adjacency pairs), and this directly implies that using conversation as a task in language testing can create considerable issues if 'interaction' is considered one of the constructs that the test is intended to measure. This point will be returned to later in the paper.

### *2.3 A definition of speaking for the KEPT context*

For the purposes of this paper, Fulcher's (2003:23) definition of speaking will be employed as a starting point, where "speaking is the verbal use of language to communicate with others". To this we can add that speaking in a conversation, as a language use situation, is governed by the grammatical, phonological, pragmatic, and sociolinguistic rules of the language given this context of use. The challenge is to consider and define which aspects of speaking this idea of speaking focuses on. First however, the context of the test will be introduced.

## 3.0 The KEPT: Context and description

### *3.1 The speaking test: The assessment context*

In the speaking test, candidates are seated facing each other in groups of four (or three in situations where one candidate is absent) and all given the same prompt which consists of a short text discussing a simple everyday topic (such as friends, family, food etc) followed by a series of simple questions for discussion which avoid specific knowledge of any subject and are written with the purpose of giving participants of all levels the opportunity to contribute to the discussion. Candidates are not scheduled with classmates or scheduled to be rated by their current teachers in order to counter as much as possible any effects due to acquaintanceship or raters allowing prior knowledge of the candidates and/or perceived ideas about their level of proficiency to influence ratings given. However, Van Moere (2006: 418) notes that the steps taken to integrate groups of candidates into groups of differing combinations of number, gender, department, year, and proficiency creates issues in that these factors could introduce construct irrelevant variation into the scores of the participants.

There are two raters in the room who do not interact with each other and individually rate all four candidates; they observe the discussion and do not interact with the students during the discussion except in exceptional circumstances where a candidate does not participate at all or participates so little that the raters feel that the candidate has not provided a rateable sample of language. In this situation one of the raters will directly attempt to involve the candidate in the discussion by asking them to share their ideas about the topic or respond to the opinion of another candidate. Raters are asked to prompt candidates a maximum of one time, and if a candidate still does not provide a sample of language they deem rateable, a score of U (or unknown) is assigned. The length of the discussion is nominally set at 8 minutes, although this varies from 7 minutes to up to 10 minutes with raters allowing the candidates to continue until they have confidently assigned scores for all categories to all candidates. Post-test the raters do not compare or discuss their ratings with each other in order to maintain intra-rater reliability as much as possible. As Van Moere (2006) notes, although one 90 minute norming session is provided before each administration of the test to the raters, these are teachers at the institution and have other duties around administration, and as such, the level of rater training could be improved.

From 1989 to 1996 a holistic rating scale was used for scoring candidate performance, and an analytic scoring rubric was introduced in 1997, and was used with some minor revisions since then until 2006. In the analytic scoring rubric there are five ‘bands’; pronunciation, fluency, grammar, vocabulary, and communicative strategies. These were broken into 4 ‘levels’ 0 – 4, although the use of half points means that in reality it was a 9 interval scale (an example of the previous scoring rubric is included in Appendix 1).

(For more information related to the goals of KEPT see Van Moere and Johnson, 2002; for more about the speaking section of the KEPT, including further discussion on recent validation efforts, see Kobayashi, Johnson, &Van Moere, 2005, and Van Moere, 2006; and for an example of past rating scales and a discussion on the difficulty of task prompts, see Bonk and Ockey, 2003).

## 4.0 Issues for consideration in defining the construct

### *4.1 Interaction*

The strongest challenge to the group oral concerns whether the test measures elements of individual's communicative ability irrespective of the other candidates' abilities. Any validity argument for a group oral test must show that construct irrelevant variation is not a significant factor in determining scores, and this aspect has been a strong focus of much KEPT related research, where studies have focused on aspect such as shyness (Bonk and Van Moere, 2004), and talkativeness (Van Moere and Kobayashi, 2004). As the current KEPT rating scales include 'communicative strategies' and the test task focuses on the communicative aspects of using language in a group discussion, the issue here concerns the extent to which interaction should be viewed as a socially derived and co-constructed phenomenon versus viewing interactional ability as a cognitive skill residing within an individual. The challenge is how the score of a candidate may change depending on the other candidates that they take the test with and the extent to which a performance that a candidate displays reflects their 'individual' performance ability. How can the construct of interest (the individual's ability to function effectively in a group discussion) and potential construct irrelevant variation (how good the other participants are) be separated? And, how can judgments be made by raters about individuals' 'conversation skill'? This point was argued by McNamara (1995, 1997) who identifies the issue: Can a demarcation be made between the candidate' s individual communicative ability if we accept that interaction is socially derived and context specific? (For more, see Kramsch, 1986). Swain (2001), examining student dialogue as a method of content specification and validation, develops this argument further, arguing that the implications of accepting a neo-Vygostkian socio-cognitive perspective leaves language testers facing the idea that interaction is a joint achievement, not individual performance.

In the literature a number of studies have discussed the issue of jointly constructed meaning in paired or group oral tests with Brown (2003) suggesting 'sympathetic' interlocutors lead to higher scores. O'Sullivan (2002), from a study on test taker familiarity, suggests that candidates varied their language when talking with familiar or relatively unfamiliar interlocutors and specifically mentions a significant effect on the judgment of the performance when Japanese female students (approximately 70% of the KEPT test population) engage in interactions (one-to-one in the O' Sullivan study) with friends and strangers.

*4.2 The performance of the current bands*

The 2006 KEPT speaking test-scoring scales have five bands: pronunciation, fluency, grammar, vocabulary, and communication strategies (see appendix 1). Underhill (1987), uses the same categories in a scoring rubric for oral assessment, where the 'communicative strategy' band of the scoring scales seem to indicate a narrow definition of strategic competence more related to Canale and Swain's (1980) interpretation of this aspect as 'coping' when a deficiency in the linguistic resources of the speaker affect their ability to communicate.

O'Donnell, Thompson, & Park, (2006) carried out a model comparison analysis on the different oral rating bands used in 2006 in order to gain some information about the performance of the bands to identify which bands need reevaluation. The results of the model comparison analysis suggested some over-correlation between the constructs, suggesting that the fluency and pronunciation bands and grammar and vocabulary bands appeared to be tapping the same constructs respectively, as the removal of up to two bands (pronunciation and vocabulary) would not significantly alter the amount of information being generated about test takers. Although this analysis was purely statistical, it highlighted two areas of concern for construct definition: to what extent are pronunciation and fluency, and grammar and vocabulary, different constructs for this assessment context?

Models of oral proficiency such as Bachman's CLA (1990), and Bachman and Palmer's (1996) model of language use, support the idea of the linguistic resources of grammar and vocabulary being more closely associated. O'Donnell, Thompson, & Park (2006), using retrospective verbal think aloud protocols to investigate the manner in which raters were arriving at the judgments of the candidates ability in the KEPT test, found preliminary support from raters to support this idea as raters complained of too many bands and difficulty in distinguishing between candidates 'grammar' and 'vocabulary' ability. This suggests that rather than interpreting the oral construct in terms of grammatical and vocabulary ability, it may be more useful to consider this both as two aspects of one overall lexico-grammatical construct.

## 5.0 Proposed test specifications for the KEPT oral

### *5.1 Towards construct definition*

As Luoma (2004:118) explains, the purpose of writing test specifications is to provide a "detailed, contextualized definition of the construct". This paper is an attempt to do just that. Luoma suggests first describing the assessment context and characteristics of the assessment procedures, before considering the nature of speaking for the test candidates, and the nature of the task that will be employed for gaining the language samples to be rated. Until now, this paper has attempted to provide insights about these aspects. The next step is one of construct definition.

Here Luoma poses three questions:

- Which models or approaches to defining language ability and speaking are relevant for this test?
- Which aspects of these models are relevant for the particular test? How are they covered in the tasks and the rating procedures?
- Which aspects are not so relevant?

*5.2 Which models or approaches are relevant?*

As Luoma (2004), and Kim (2006) explain, Bachman's (1990) Communicative Language Ability (CLA) and Bachman and Palmer's (1996) model of Language Ability are generally considered the most comprehensive attempt at a general model of language ability (it should be noted that for the purposes of the discussion in this paper, the term CLA model will be used to refer to both the original 1990 Bachman CLA and the 1996 Bachman and Palmer model of language ability).

Bachman and Palmer divide language ability into two parts; a static "domain of information in memory" (1996: 67) which compromises their 'language knowledge' which is mediated in different language use situations by "a set of metacognitive strategies components, or strategies… that provide a cognitive management function in language use" (p.70), which Bachman and Palmer refer to as 'strategic competence'. Here a clear difference between previous models of communicative language ability can be drawn, as the CLA moved strategic competence squarely to the center of communicative competence, where individual's strategic competence is the mediator between their linguistic and world knowledge in the context of the language use situation. For more on the CLA see Bachman (1990), and Bachman & Palmer, (1996).

Fulcher's Speaking Framework (2003) is specifically focused on speaking and is presented in terms similar to the CLA model. This provides a clear movement from a general theoretical model of communicative language ability, to a 'skill' based or what could be called a 'specific channel of language use activity' level. Fulcher' s suggested framework comprises five categories: language competence, strategic capacity, textual knowledge, pragmatic knowledge, and sociolinguistic knowledge. In Fulcher's model, language competence is divided into three areas: 'Phonology' comprising pronunciation, stress, and intonation; 'Accuracy' comprising syntax, vocabulary, and cohesion; and 'Fluency' comprising hesitation, repetition, re-selecting appropriate words, re-structuring sentences, and cohesion. For Fulcher, the concepts of fluency and accuracy seemed strongly linked as "accuracy and fluency are associated with automaticity of performance and the impact this has on the ability of the learner to understand" (2003:31).

*5.3 Consideration of the models in the KEPT assessment situation*

Fulcher's division of 'textual knowledge' from 'language competence' suggests that to some extent this knowledge of the 'structure of conversations" (2003:34) Fulcher views as a separate area of 'speaking' and conversational structural knowledge, as opposed to general language ability. Alternatively it could be considered as separate because of the 'interactional' aspect of conversation structure as more communicatively focused in language use in spoken discourse, versus the obviously psycholinguistic characteristics of phonology, accuracy, and cohesion. The implications of this division, along with Fulcher's reluctance to employ more than an individually focused view of strategic language use in conversation, suggests a narrow view of 'interactional skill' at the current time in response to the criticisms of those such as McNamara (1995, 1997) and Swain (2001) about the idea that interaction is a joint achievement, not individual performance. The Fulcher framework provides an outline of aspects that can be considered to be observable performances of individuals, rather than outcomes stemming from the interaction, which is important to remember for the KEPT.

The 2006 communicative strategies band emphasizes ‘interaction, confidence, conversational awareness’ (see appendix 1), with a level two score (considered the criterion level) being described as ‘Responds to others without long pauses to maintain interaction; shows agreement or disagreement to others’ opinions’. When examined against the speaking framework offered by Fulcher, this band could be seen to be attempting to measure the students textual knowledge through their understanding of turn taking (responding to others to maintain the interaction), and their sociolinguistic knowledge (participation in the conversation and willingness to share ideas). Nevertheless, the idea of ‘interactional competence’ remains an underlying idea behind this construct, best described by Fulcher (2003:44) who explains a common definition of the term as “how speakers structure speech, its sequential organization and turn-taking rules, sometimes including strategies”. For the purposes of the KEPT oral construct, this definition seems to best fit what the current bands are attempting to measure: individual’s ability to function and participate effectively in a group discussion. The criticisms of this idea (McNamara, 1995, 1997; Swain, 2001) remain, and this aspect of the construct of the test seems in the most need of more research.

O’Donnell, Thompson, & Park’s (2006) model comparison analysis indicated the importance of fluency as a construct for the KEPT oral, and therefore it is important to consider and discuss the issues surrounding fluency in working towards some definitional clarity of the construct for test specification purposes. Speaking speed, use of fillers, and degree of automatization are the primary aspects which raters are asked to consider. This indicates that the way the fluency construct can be interpreted is very close to the idea expressed both in Fulcher's framework (2003), and in Fulcher's 'fluency descriptors' (2003:250-252) where degree of automaticity (and subsequently speaking speed) is indicative of the extent to which candidates appear to have 'lexicalized' language as chunks which are readily available for use versus the idea that they are forced to 'syntactically create' utterances based on their understanding of formal rules of grammar, collocation, and vocabulary. In other words, the question is the extent to which the candidates can express themselves smoothly in real time, indicating that they are able to control their lexico-grammatical choices under the processing time constraints of the test in such a way as to present more or less grammatically correct utterances in context. The inference is that this indicates internalization has gone on which allows them to concentrate more on semantic meaning they wish to convey – i.e. language has been acquired and is available for use.

The 'Pronunciation', 'Grammar', and 'Vocabulary' bands make up the remaining 3 rating scales in 2006, and comparing these against the CLA model and Fulcher' s speaking framework show that the test is heavily focused on aspects of what Bachman and Palmer (1996) refer to as 'Organizational knowledge'. One aspect that is not covered in the CLA or Fulcher's framework (being a relatively small concept) is a clearer consideration of the issues surrounding range of language (vocabulary and syntactic) used versus correctness of language used. Some have noted the relatively narrowly definition of grammatical ability in the CLA model (Kim, 2006, citing Purpura, 2004), arguing that the concept of grammatical ability should in itself be more communicatively oriented and include the idea of intended meaning and grammar in use rather than limiting this to 'functions'. The issue of 'accuracy', and 'complexity', and how users maximize their resources may be considered an area of strategic competence in the CLA, but this is an issue to remember when attempting to operationalize the construct in the rating scales, as this aspect seems to be of concern to raters (O'Donnell et al, 2006). Thus it is important to take an expanded view of the CLA to include the pragmatic (meaning focused) aspects of language ability as defined in the CLA, so that grammar and vocabulary are seen as not only formally oriented, but also semantically oriented, including the intended meaning of the utterance. Furthermore, The CLA's concepts of sociolinguistic knowledge may be operationalized through the use of idiomatic expressions or terms and therefore may be included in an expanded idea of grammatical and vocabulary ability which includes formal understanding and performance, combined with appropriate usage in conveying intended meaning naturally and appropriately. Therefore, aside from the issue of accuracy and complexity, the CLA model seems to fit the structural aspects of language knowledge well.

## 6.0 Towards definition of the construct

### *6.1 Descriptive definition of the construct*

The aim of this test is to assess the examinees' general oral proficiency within the context of a group discussion. As a test of spoken interaction in a foreign language (English) which is designed to reflect the values of the institution (Kanda University of International Studies), communicative language ability as defined by Bachman (1990) and Bachman and Palmer (1996) best explains the view of language ability to be operationalized in this assessment, where individual's strategic competence mediates their language ability, topical knowledge, affect, and personal characteristics with the language use situation. The examinees' oral proficiency is seen to be reflected in their ability to communicate smoothly and naturally with the other examinee interlocutors, expressing their opinions and ideas, and involving other examinees in the discussion.

In terms of pronunciation and fluency, a good performance would show the examinee's ability to express them self clearly and comprehensibly, showing little or no first language interference and an ability to integrate supra-segmental and intonation aspects of English pronunciation into their speech such as word and sentences stress, rhythm and weak forms, and intonation contours. Good performances would also demonstrate the examinees ability to form utterances in real time smoothly, indicating high levels of planning ability, automaticity, and lexicalized knowledge, where the examinee would demonstrate the ability to speak in extended discourse, integrating fillers in a natural manner during pauses, with any slips quickly repaired and reformulated. The performance would also show the examinee's ability to use their linguistic resources (grammar and vocabulary) effectively to express their ideas by structuring information clearly and relevantly as part of a coherent discourse where the examinees would understand the other interlocutor's turns and fit their own turns accordingly. They would seldom misunderstand their interlocutors and may grade their language when dealing students of lower level. Excellent performances would also reflect the ability of the examinee to accurately use a range of vocabulary and structures with complete appropriacy and these examinees would be able to confidently and naturally interact and participate in the discussion.

A poor performance would be typified by little or no evidence of language ability, with little or no attempt to participate in the discussion and pronunciation strongly influenced by the first language which may strongly hinder communication. In a poor performance, the examinee will often show signs of not having comprehended others' utterances, will have difficulty recognizing appropriate turn offers, and would have trouble responding to direct questions with responses of more than a few words expressed in limited utterances. Frequent unnatural pauses and recasts would be present with avoidance and/or abandonment strategies possible, although these may be sometimes difficult to observe. In a bad performance, examinees will not be able to coherently structure even relatively simple transactional discourse, and even high level interlocutors may appear to have difficulty understanding the meaning of their utterances at times where the examinee has insufficient grammatical or lexical knowledge to explain ideas in any extended discourse. Although they will be able to respond to simple direct questions, they would not be able to build on these turns and create a continuing conversation or coherent discourse.

Fluency is viewed as an important aspect of this test, reflecting the content of the curriculum, which is generally focused more on building speaking speed, confidence, and automaticity rather than accuracy and correct formal usage and application of grammatical rules. Subsequently, for the purposes of this test, fluency is operationalized specifically in the extent to which an examinee displays ability to speak in extended discourse (utterances consisting of more than three clauses), their rate of speaking speed, and the extent to which they integrate fillers and other devices into their speech in an appropriate manner to hold the floor, buy time while considering ideas, or otherwise signal understanding of the structure of conversations.

Knowledge of words, phrases, and grammatical structures are also important with vocabulary and grammatical ability being viewed as one construct (lexico-grammatical) which is operationalized in terms of the range of vocabulary and structures used, along with the accuracy and appropriacy of the structures and vocabulary items used. A key idea in this construct is a focus on meaning, and to this end the construct can be further described as the examinee's ability to correctly form utterances, and their ability to integrate (or attempt to integrate) a variety of grammatical structures and lexical items in their language use appropriately to convey meaning effectively.

Finally, interaction and participation are considered important factors in this test, reflecting examinees' ability to understand conversation function commonly experienced in in-class activities such as turn-taking, responding to others, offering opinions, providing feedback, and employing strategies to repair misunderstandings through paraphrase and/or examples, and asking for or providing clarification. The way this aspect is being operationalized can be considered a combination of two aspects of Bachman and Palmer's CLA  concerning individuals' textual knowledge and strategic competence. However, it is recognized that this idea of 'interactional competence' creates potential for construct irrelevant variation as the issue of the co-construction of meaning in communication, especially in conversations between equal rights holders, potentially creates a situation where identifying an individual's true 'communicative skill' is impossible due to factors outside their control, such as the groupings of examinees.

## 7.0 Limitations of the research and suggestions for further research

### *7.1 Limitations of the research*

In an investigation into the validity and reliability of different large scale language proficiency tests, Chauhoub-Deville and Turner (2000, p.536) state that "in test development, construct delineation is a prerequisite to rendering meaningful scores" and cite Anastasi (1986: 3, cited in Chauhoub-Deville & Turner, 2000: 536), who explains the need to build validity into the test development process from the beginning where "The validation process begins with the formulation of detailed trait or construct definitions, derived from psychological (communicative) theory, prior research, or systematic observation and analyses of the relevant behavior domain". This is echoed by Fulcher (1996: 170), who rightly notes in a critique of validation research conducted on the American Council on Teaching of Foreign Languages (ACTFL) rating scales, "it is vitally important" to "consider construct validity at the test development stage of the process, rather than as a post hoc activity". It thus should be noted here that the purpose of this paper is not to prescribe a set of test specifications for the oral subsection of the KEPT, but rather serve as a discussion document for the KEPT committee in an ongoing project to reinterpret the oral constructs of the test and rating scales in the development of new test specifications. Fulcher's point is important though in highlighting the primary limitation of this paper, in that while it draws together some historical information related to the KEPT, outlines research into the KEPT until this point, discusses the relationship of other empirical and theoretical work in the field of language testing, and uses these as the basis for an interpretation of how oral proficiency may be interpreted in this testing situation, the judgments made are not supported by empirically derived information. As a result, the important step that is required is the integration of a data-based approach to rating scale and test construct design to inform the further development of test specifications and rating scales. It is hoped that this paper provides some of the necessary background to allow the process to proceed in a somewhat iterative manner, with empirical data and theory being synthesized in the development of a comprehensive set of test specifications. With that in mind, a number of suggestions for further research related to the KEPT oral can be made.

### *7.2 Proposed Rating bands*

For the purposes of giving some insight as to how the test specifications and construct outlined in section 6 of this paper could be interpreted in a set of rating bands, a draft version of revised rating rubric is attached in appendix 2.0. It should be noted that these draft rating bands account for the results of the O'Donnell et al. (2006) model comparison data and think-aloud protocol research, along with Batty' s (2006) vocabulary depth correlational research which supported the idea that the vocabulary band was working ineffectually. The result is a collapsed 'Accuracy and Complexity' band which follows Skehan's (1997) operationalization of 'grammatical knowledge' which includes vocabulary use in a lexico-grammatical construct which also accounts for complexity of language use, rather than only focusing on form. Furthermore, 'communicative strategies' has been reinterpreted as 'conversational awareness' and strategies which emphasize a number of conversational functions that the students encounter in the University curriculum (offering opinions, responding etc), conversational organization knowledge in turn-taking and discourse management, along with Fulcher's communicative strategies. Fluency is primarily seen as a measure of automaticity of linguistic resources and for this reason is considered a different construct to pronunciation, which remains relatively unchanged and focused on the sound system and articulation of speech.

### *7.3 Future steps for validating and further refining the construct*

As McNamara (1997: 448) explains, the fundamental issue with validation is "how you can defend the conclusions you have reached about the person you are assessing when they are based on the limited sample of his/her performance that is available through the test setting". This is the challenge faced by all test creators, and one answer is suggested by Chauhoub-Deville and Turner (2000: 525), who explain:

> The validation process is not a one-time activity but an ongoing process. Validation research emphasizes an ongoing and a systematic research agenda that documents the properties and interpretations of test scores and provides evidence to support their use as specified in the test purpose.

The validation process includes gathering evidence regarding the relevance and the representativeness of the content covered from the specified construct domain (Messick, 1989, 1996). Additionally, validation emphasizes theoretical arguments and empirical evidence to support test score interpretation and use. This view of validation as a collection of evidence to support and justify the inferences made about the language ability of the test takers (Fulcher, 2003; Luoma, 2004) points to a requirement for more KEPT specific research. While this paper has outlined a number of studies investigating the KEPT oral test which have provided evidence to support a validity argument for the test, the relatively low inter-rater reliability coefficient of 0.74 in the Van Moere study (Van Moere, 2006) suggests further development is required in the test and Van Moere's (2006:436) warning about the danger of increasing the stakes of the test suggest some areas where more empirical data is required:

Cut score decisions based on a single test score would necessarily have to take into account a wide margin for error, giving candidates the benefit of the doubt for being assessed in a group that may have restricted their performance, portrayed them in a bad light, or otherwise affected their ability.

The recurring issue of interaction in group oral testing which has surfaced throughout this paper remains the area that information is needed about most and the test construct as defined in this paper remains open to criticism without empirical evidence to support its continued inclusion. This may be the single biggest issue in the construct reevaluation process and it appears that the validity argument for this type of test will require extensive data to account for the potential different forms of variance that the format allows. Furthermore, generating data on the extent to which bias occurs in the speaking test is another aspect of the oral test which could provide further evidence to support a validity argument for the test, or raise issues for the test development team to deal with. Differential item function (DIF) research could provide valuable insights and information into potential rater bias in the KEPT speaking test.

Finally, Upshur and Turner (1999) argue that rating scales should not only be population specific but also task specific, a suggestion echoed Chaulhoub-Deville (1995) and Fulcher (1996) who emphasizes not only task specificity but a data-based approach to rating scale design. Thus, the next step in continuing the re-evaluation of the KEPT oral test construct and rating scales appears clear: gathering data to empirically consider aspects of the task, including the types of discourse that alternative task types generate in group oral tests, as well as aspects of language generated from the task to validate, or provide the basis for further reevaluation of the test construct and scoring bands.

## 8.0 Conclusion

This paper has constituted an attempt at highlighting and critically examining some of the aspects of the construct of the speaking subsection of an in-house English language proficiency test from a University in Japan. Some initial discussion was carried out concerning the nature of the speaking 'skill', with the conclusion that for this assessment context, a more interactionalist approach based on Bachman and Palmer's CLA model (1996) was appropriate.

A recurring theme throughout this paper has been the issue of interaction in communication, and responding to the challenge of the co-constructed view of interaction is the greatest challenge facing the KEPT oral if that particular aspect of the proposed test specifications is to be accepted. As Bonk (2001:83) explained, what is essential for the KEPT is defining what elements of proficiency are essential "for the kinds of interpretations that we would like to make about our students". Given the strong support for the format of the test within the institution and positive washback effects on the curriculum in stimulating conversational ability and communication focused speaking, facing the challenge of defining interactive competency is critical, and relating this idea to observable performances that reflect the examinees' ability to effectively participate and use language in conversation is a major challenge for the KEPT committee. This paper has attempted to start this process, and in the proposed set of specifications, modeled on Bachman (1990) and Bachman and Palmer's (1996) CLA model along with Fulcher's (2003) framework for describing speaking, fluency and communicative skill are critical components. It is hoped that this paper serves as a discussion document that generates further examination of the test construct and requirements of the task. The ongoing need for more validation efforts with respect to the construct of the test, and particularly with respect to the place of interactive 'competence' within the test construct, remain the greatest challenge for building the body of validation evidence for this test.

## 10.0 References

Bachman, L.F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L., Lynch, B. & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing 12,* 238–58.

Bachman, L.F., & Palmer, A.S. (1996). *Language Testing in Practice.* Oxford: Oxford University Press.

Batty, A. (2006). Vocabulary Depth in written and oral assessment. Paper presented at the *JALT International Conference,* Kitakyushu, Japan.

Bonk, W. (2001). Predicting paper-and-pencil TOEFL scores from KEPT data. *Studies in Linguistics and Language Education of the Research Institute of Language Studies and Language Education, Kanda University of International Studies, 11,* 163 – 224.

Bonk, W.J., & Ockey, G.J. (2003). A many facet rasch analysis of the second language group oral discussion task. *Language Testing, 20*(1) 89-110.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*(1), 1-15.

Canale, M. (1983). On some dimensions of language proficiency. In J.W. Oller (Ed.). *Issues in language testing research.* Rowley, MA: Newbury House.

Canale, M., & Swain, M., (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1,* 1- 47.

Chaulhoub-Deville, M. (1993). *Performance assessment and the components of the oral construct across different tasks and rater groups.* (Tech. Rep. No. 143). Columbus, Ohio: Ohio State University. ERIC No. 360 830

Chaulhoub-Deville, M., & Turner, C.E. (2000). What to look for in ESL admission tests: Cambridge certificate exams, IELTS, and TOEFL. *System, 28,* 523 -539.

Fulcher, G., (1996). Testing tasks: issues in task design and the group oral. *Language Testing, 13,* 23- 51.

Fulcher, G. (2003). *Testing Second Language Speaking.* Harlow: Pearson Education.

Goffman, E. (1976). Replies and responses. *Language in Society, 5,* 254-313.

Hymes, D. (1972). On communicative competence. in J.B Pride and J. Holmes (Eds.), *Sociolinguistics.* Harmondsworth: Penguin, pp. 269-293.

Jones, E.E., & Gerard, H.B. (1967). *Foundations of social psychology.* New York: Wiley.

Kim, H. (2006). Providing validity evidence for a speaking test using facets. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 6*(1). New York: Columbia University.

Kobayashi, M., Johnson, K., &Van Moere, A., (2005). Effects of quantity and quality of students' output in group oral tests. *Studies in Linguistics and Language Teaching of the Research Institute of Language Studies and Language Education, Kanda University of International Studies, 16,* 275-295.

Kormos, J. (1999). Simulating conversations in oral-proficiency assessment: A conversation analysis of role plays and non-scripted interviews in language exams. *Language Testing, 16* (2), 163-188.

Kress, G., & Fowler, R. (1979). Interviews. In R. Fowler, R. Hodge, G. Kress, and T. Trew (Eds.), *Talking and testing: discourse approaches to the assessment of oral proficiency.* Amsterdam: John Benjamins, 239-67.

Kramsch (1986). From language proficiency to interactional competence. *Modern Language Journal, 70*(4), 366-72.

Luoma, S. (2004). *Assessing Speaking.* Cambridge: CUP.

McNamara, T.F. (1995). Modelling performance: Opening Pandora's box. *Applied*

*Linguistics, 16*(2): 159-179.

McNamara, T.F. (1997). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics, 18*(4), 446-466.

Messick, S., (1989). Validity. In Linn, R.R. (Ed.), Educational Measurement, *3rd Edition.* New York: American Council on Education/Macmillan, pp. 13-103.

Messick, S., (1996). Validity and washback in language testing. *Language Testing, 13,* 241-256.

Ockey, G.J. (2001). The group oral and the oral interview. Paper presented at the *Teachers of English to Speakers of Other Languages (TESOL) Conference,* St Louis, Missouri, USA.

O'Donnell, D.O., Thompson, G.R., & Park, S. (2006) Revisiting assessment criteria in a speaking test. Paper presented at the *Japan Association of Language Teachers (JALT) conference,* Kitakyushu, Japan.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing, 19*(3), 277–295

Purpura, J. (2004). *Assessing second language grammar ability.* Cambridge: CUP.

Sacks, H., Schegloff, E.A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking in conversation. *Language, 50,* 696-735.

Silverman, D. (1973). Interview talk: Bringing off a research instrument. *Sociology, 7,* 31-48.

Skehan, P. (1998). *A cognitive approach to Language Learning.* Oxford: OUP.

Skehan, P., & Foster, P. (1997). The influence of planning and post-task activities on accuracy and complexity in task-based learning. *Language Teaching Research, 1,* 185-211.

Swain, M. (2001). Examining dialogue: Another approach to content specification validating inferences drawn from test scores. *Language Testing, 18*(3), 275 – 302.

Upshur, J. and Turner, C. (1999). Systematic effects in the rating of second language speaking ability: Test method and learner discourse. *Language Testing, 16,* 82–111.

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing, 23*(4) 311-440.

Van Moere, A., & Johnson, F.C. (2002). Communicative assessment in a personal curriculum at Kanda University of International Studies. In Mackenzie, A. and Newfields, T. (Eds.), *Curriculum innovation, testing and evaluation.* Tokyo: The Japan Association for Language teaching (JALT), College and University Educators (CUE) and Testing and Evaluation (TEVAL) Special Interest groups (SIGS), 155-62.

Van Moere, A. and Kobayashi, M. (2004). Group oral testing: Does amount of output affect scores? Paper presented at the Language Testing Forum.

## 11.0 Appendices

**Appendix 1**

### *Oral Descriptor Bands 2005*

| | Pronunciation<br><br>Think about:<br>• pronunciation<br>• intonation<br>• word blending | Fluency<br><br>Think about:<br>• automatization<br>•fillers<br>• speaking speed | Grammar<br><br>Think about:<br>• use of morphology<br>• complexity of syntax (rembedded clauses, parallel structures, connectors) | Vocabulary<br><br>Think about:<br>• range of vocab | Communicative skills/strategies<br><br>Think about:<br>• interaction<br>• confidence<br>• conversational awareness |
|---|---|---|---|---|---|
| **0**<br>**.5** | Very heavy accent, uses Japanese katakana-like phonology and rhythm; words are not blended together | Fragments of speech that are so halting that conversation is not really possible; nss would not think person had virtually no English | Does not use any discernable grammatical morphology | Shows knowledge of only the simplest words and phrases taught in junior high school or beginning high school | Shows no awareness of other speakers; may speak, but not in a conversation-like way |
| **1.0**<br>**1.5** | Somewhat katakana-like pronunciation; does not blend words together, they are pronounced in isolation | Slow strained speech, constant groping for words and long unnatural pauses; communication with a ns would be difficult | Doesn't have enough grammar to express an opinion clearly; makes frequent errors; no attempt at complex grammar | Lexis not adequate for task, cannot express opinion properly with the limited words used | Does not initiate interaction, produces monologue only; shows some turn-taking, may say, "i agree with you," but not relate ideas in explanation; too nervous to interact effectively |
| **2.0**<br>**2.5** | May not have mastered some difficult sounds of English, but would be mostly understandable to a naïve NS; makes some attempts to blend words | Speech is hesitant; some groping for words and unfilled spaces are present but generally don't impede communication completely | Relies mostly on simple (but appropriate) grammar, has enough morphosyntax to express meaning, complex grammar is attempted but may be inaccurate | Generally has enough lexis for expressing some opinion but does not demonstrate any particular knowledge of vocabulary | Responds to others without long pauses to maintain interaction; shows agreement or disagreement to others' opinions |
| **3.0**<br>**3.5** | Pronunciation is good but has still not mastered the sound system of English; accent does not interfere with comprehension; can blend words | May use some fillers, rarely gropes for words but speech may still not be quick | Shows ability to use some complex grammar, may make errors but they are only in late-acquired grammar | Shows some evidence of some advanced vocabulary | Generally confident, responds appropriately to others opinions, shows ability to negotiate meaning quickly and relatively naturally |
| **4** | Speaks with excellent pronunciation and intonation; has practically mastered the sound system of English | Excellent fluency, uses fillers effectively, shows ability to speak quickly in short bursts | Uses both simple and complex grammar effectively; may make occasional errors but they are only in late-acquired grammar | Shows evidence of a wide range of vocabulary knowledge | Confident and natural, asks others to expand on views, shows how own and others' ideas are related, interacts smoothly |

Note: If a student shows she is consistently fulfilling the criteria being tested, she receives the score at the bottom of the box; if she *sometimes* achieves the expected level, but sometimes slips to a lower criteria, she is given the higher score in the box.

If a student did not speak enough for you to reliably assign a score for a category, see if you can get them to speak more. If they don't oblige, assign them U as a score for that category.

## Appendix 2

| | **Pronunciation** (1/2 weight)<br>Think about:<br>**• Phonemic level: individual sounds**<br>**• Word level: stress and weak forms**<br>**• Sentence Level: ability to 'blend' or link sound within or between words.**<br>**• Sentence stress, rhythm, and intonation**<br>**• Degree of first language phonological interference** | **Fluency**<br>Think about:<br>**• Automatization: ability to formulate utterances quickly and speak in extended discourse smoothly (indicating lexicalized knowledge)**<br>**• Speaking speed**<br>**• Hesitations and pausing (unnatural groping, considering next expression, pre-planning, false starts, reformulations)**<br>**• Misunderstandings**<br>this band adapted from Fulcher (1996) | **Accuracy and Complexity**<br>Think about:<br>**• Correct grammatical formation of utterances and suitability of vocabulary used**<br>**• Displaying ability to use (or attempting to use) different grammatical structures and vocabulary suitably in context.**<br>**• Collocations and correct word choice** | **Conversational awareness and strategies**<br>Think about:<br>**• Communicative strategies (avoidance of unknown words, abandonment, approximation, word creativity, circumlocution)**<br>**• Participation and smoothness of interaction (turn-taking, responding to others, asking questions and introducing new gambits, paraphrasing, hedging)**<br>**• Turn-taking and discourse awareness** |
|---|---|---|---|---|
| **0**<br>**.5** | Very heavy first language interference (accent),<br>Relies on Japanese 'katakana'-like phonology and rhythm.<br>So little control of the individual sounds as to make understanding almost impossible | Fragments of speech<br>Halting, often incomprehensible | Does not use any discernable grammatical morphology or vocabulary with communicative force<br>Some simple vocabulary may be used in isolation but in general appears unable to share simple ideas | no awareness of other speakers<br>appears to be unable to follow the conversation<br>may speak, but only in random utterances<br>makes not attempt to join the conversation |
| **1.0**<br>**Poor**<br>**1.5** | Somewhat 'katakana'-like pronunciation (high first language pronunciation interference)<br>Little control of individual sounds Little or no understanding of connected speech patterns<br>Does not blend words together (pronounced in isolation) | Utterances fragmented or incomplete<br>Frequent extended pauses while completing utterances and slow speech<br>Often appears to misunderstand interlocutors<br>Often appears unable to respond to questions | Appears not to have mastered even simple grammatical forms and frequent errors make meaning difficult to understand<br>Very little control or range of vocabulary which heavily impedes communicative force<br>No attempt at using complex grammar and struggles with even simple collocation pairs | Has strong difficulty following or joining the discussion where appropriate unless directly asked<br>May show simple agreement or disagreement, but seems to lack ability to provide simple feedback (Yes, hmm)<br>May abandon utterances completely and leave the conversation 'hanging'<br>Does not continue or develop turns, participating randomly |
| **2.0**<br>**Fair**<br>**2.5** | Has mastered most simple individual sounds of English<br>Limited ability to integrate word and sentence stress, weak forms, and intonation in connected speech<br>May commonly use word stress incorrectly (e.g. verb versus noun forms) | Usually able to complete utterances, but still slow production<br>Frequent short 'mini' pauses during utterances due to slow production<br>Can usually respond to questions, but may require considerable planning time<br>May attempt to use fillers | Tends to rely on a small range of relatively simple grammar and vocabulary (*1)<br>Grammar used is mostly formed and used correctly to communicate meaning<br>When lower frequency vocabulary or more difficult grammar is attempted, shows a lack of control or inability to formulate correctly | Seems able to follow conversation and joins the conversation on their own (by offering comments of suggestions)<br>Responds to others without long pauses to maintain interaction<br>Shows meaningful agreement or disagreement to others' opinions (elaborated assent / dissent)<br>Changes in turns (especially receiving) may be delayed or are not smooth<br>May provide some limited feedback to speakers when listening |
| **3.0**<br>**Very**<br>**good**<br>**3.5** | Has mastered individual sound system of English to a high level<br>Accent does not interfere with comprehension<br>Appears to have mostly mastered simple word stress rules (e.g. verb versus noun forms)<br>Demonstrates some ability to 'blend' words and control intonation in connected speech | Pauses are more natural, indicating word choice more than grammatical formulation<br>May integrate fillers during time when they require pauses for planning or reformulation<br>Able to smoothly respond to questions with few misunderstandings<br>Displays some ability to speak in extended discourse | May demonstrate ability to use a range of grammatical forms but often makes mistakes<br>Makes very few mistakes with a somewhat limited range of grammatical forms<br>Makes an effort to integrate a range of vocabulary<br>Shows some ability to use vocabulary suitable in terms of collocation or idiom, but may lack control at times | Smoothly joins the discussion without long pauses between speakers<br>Noticeably back-channeling and providing feedback during conversation (hmm, yeah) but may not display a range of fillers or ability to use them for different functions<br>Responds appropriately to others opinions, and attempts to include others in the interaction appropriately<br>Introduces new gambits and attempts to negotiate meaning when required (paraphrase, clarification) |
| **4**<br>**Excellent** | Has practically mastered the individual sound system of English<br>Speaks with excellent control of word and sentence stress<br>Intonation in speech is controlled with no first language interference in stress patterns, intonation, or rhythm | Almost never hesitates and pauses that do occur don't reflect lack of language ability<br>Some natural planning between utterances<br>Fillers or other devices are used naturally to fill pauses during speech<br>Displays ability to speak in extended discourse at a natural rate and speed<br>Displays confidence in their ability to get things right, and rarely re-formulates<br>When they seem obviously aware of a grammatical slip, they reformulate and repair accordingly<br>Shows ability to speak quickly in short bursts | Uses a range of grammatical forms with high accuracy<br>May make small errors which do not impede communicative force (*2)<br>Demonstrates excellent control of vocabulary<br>Shows ability to use a range of vocabulary in a natural manner (such as idiom, collocation etc) | Completely confident introducing new gambits, involving other interlocutors, and asking others to expand on views through question or paraphrase<br>May appear to be grading language or hedging opinions, cooperatively constructing or repairing other's mistakes, or paraphrasing where appropriate to ensure others understanding<br>Noticeably provides feedback (hmm, yeah) for different functions (show listening, agreement, understanding)<br>Holds and relinquishes turns appropriately and naturally, and shows complete confidence with participation in the interaction at all times |

**IMPORTANT NOTE:** If a student did not speak enough for you to reliably assign a score for a category, see if you can get them to speak more. If they don't oblige, assign them U as a score for that category. Please use ½ scores (e.g. 2.5) when a student shows she is consistently fulfilling the criteria of the level (2), and is *sometimes* achieving aspects of the next level (3).

*1: Simple verb forms (present, past, future using will)  *2: Esp. articles

| | |
|---|---|
| <1 | Non-user |
| >2 | Exit Threshold |
| > 3 5 | Kanda expert-user |
| U | Un-gradable |