ISSN 1941-7233 (Print), ISSN 2345-024X (Online)
http://dx.doi.org/10.7220/2345-024X.16.3

ENGLISH SPEAKING SKILLS ASSESSMENT FOR GRADE 6 THAI STUDENTS: AN APPLICATION OF MULTIVARIATE GENERALIZABILITY THEORY

Daruwan Srikaew, Kamonwan Tangdhanakanond[1], Sirichai Kanjanawasee
Chulalongkorn University, Thailand

Abstract. ***Background and Purpose***. This research studies an analytic rating scale for an English speaking skills assessment designed for Grade 6 Thai students learning English as a foreign language. ***Material and Methods***. A rating scale was developed, validated, and then piloted using assessment results from 101 students attending a government school in Bangkok, Thailand. The analytic rating scale developed for the assessment is composed of 6 components: vocabulary, syntax, cohesion, pronunciation, ideational function and fluency. The reliability of the rating scale was examined with different numbers of speaking tasks and raters. Multivariate generalizability theory (G theory) was utilized for the data analysis. ***Results***. The results showed that fluency was the greatest variance component of the composite score of the analytic rating scale, followed by ideational function, cohesion, vocabulary and syntax, and pronunciation respectively. Reliability of the composite score for the speaking analytic rating scale was high (over .80). It was found that the reliability coefficients for each component would be reliable (over .80) when six or more tasks are used and the number of raters is from 6–10 and above. The dependability increased more when the number of tasks increased and when the number of raters increased. It was also found that a reliable high Phi Coefficient (over .80) could be obtained using only 6 tasks and 3 raters. ***Conclusion***. The main results were then discussed in more detail.

Keywords: English speaking skills, analytic rating scale, multivariate generalizability theory.

[1] Address for correspondence: Department of Educational Research and Psychology, Faculty of Education, Chulalongkorn University, Bangkok, Thailand. Email: tkamonwan@hotmail.com

INTRODUCTION

Assessing language proficiency is important because language is a tool for communication between people and for international linkage on different businesses. English is one of the most important languages. It is even more important at present when Thailand, as well as the other members of The Association of Southeast Asian Nations (ASEAN) countries, is at the preparation stage for implementation of the ASEAN community in 2015. English speaking skills is one of the language skills which primary school students must develop along with other skills. It is a beneficial skill for students in their learning process and in development of higher level of communication. Students' ability in communicative language should be assessed using performance tests and authentic assessments that allow students to speak, talk and perform in realistic contexts (Bachman & Palmer, 2010). Bachman and Palmer (2010) defined language ability as the ability to use language for communication which consists of (a) language competence or language knowledge and (b) language strategic competence. Language knowledge includes two broad categories: organizational knowledge and pragmatic knowledge as shown in Figure 1.

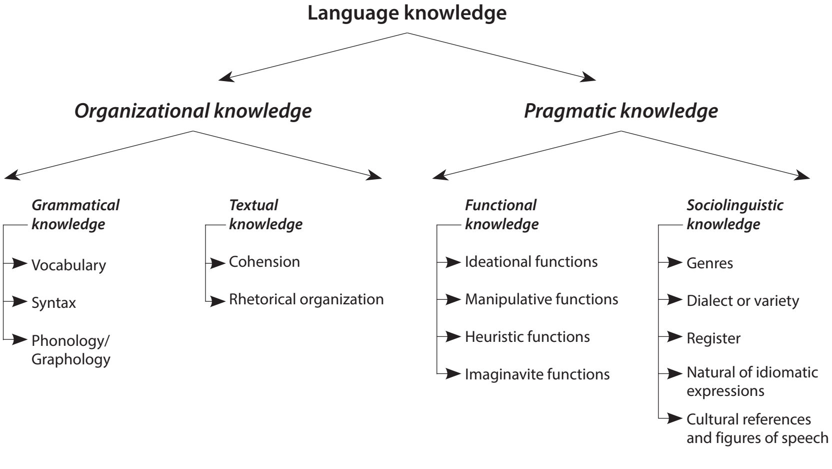

Figure 1. *Bachman and Palmer model of areas of language knowledge (2010).*

Bachman and Palmer (2010) stated that organizational knowledge consists of grammatical knowledge and textual knowledge. Grammatical knowledge consists of three components: vocabulary, syntax and phonology/graphology; whereas textual knowledge consists of two components: cohesion and rhetorical or conversational organization. As for pragmatic knowledge, it is the knowledge related to using language in practice, such as the ability to communicate and achieve the desired objectives, and using appropriate language in different situations such as functional knowledge and sociolinguistic knowledge. Speaking skill assessment involves many activities and many parts. The components associated with the process of speaking skill assessment are interrelated (Luoma, 2004) and include students, evaluator, assessing tools or tasks, and qualitative scoring scheme for speaking skill assessment.

The rating scale for English language ability has been classified by linguists in varying ways. Weigle (2002) presented two models of the rating scale for writing assessment: (a) Holistic rating scale and (b) Analytic rating scale. The Analytic rating scales have been used for more reliable assessment, since it provides clear information that is useful for higher level learning development (Elbow 1991, cited in Barton & Collins, 1997). Mostly, the scales were no more than 4 levels (Weigle, 2002). It was found that five components were assessed in the speaking skill assessment: (a) pronunciation, (b) vocabulary, (c) cohesion, (d) organization, and (e) grammar (Sawaki, 2007). During the research study of 6-11 year-old primary school students, it was found that the components that have been assessed included grammatical accuracy, fluency, scope of vocabulary, pronunciation and content (Efthymiou, 2012). The research on explanation speaking skills of primary school students (Westerveld & Moran, 2011) assessed students in four components that were verbal productivity, syntactic complexity, grammatical accuracy and verbal fluency.

In English speaking assessment, students must demonstrate their speaking ability. Because the subjectivity and bias of raters are often unavoidable in speaking assessment scoring, reliable scoring from raters, appropriate rating scales, and appropriate tasks are very important. It was found in the past that the validity of language assessment, either on speaking, reading or writing, was low, since the related variables, such as number of tasks, number of raters, scoring scales and writing mode, were confusing (Cooper, 1984; Gebril, 2009; Huot, 1990; Lee & Kantor, 2005; Schoonen, 2005, cited in Gebril, 2010). Therefore, efforts were put into developing a more reliable language assessment.

The generalizability theory or G-theory is a tool to analyze the relationship between a composite score and each part of test components including number of examinees, number of test items, raters and other sources of testing errors (Brennan, 2001; Lisa, Brian, & David, 2010). Multivariate generalizability theory (MGT) provides more information than generalizability theory because it can determine the relationship between universal score and the test components, or between the test of components and the grouping condition in order to make the highest validity of composite scores (Brennan, 2001; Lee, 2006; Webb, Shavelson, & Haertel, 2007). It also provides information about the relationship between parts of the test and helps to verify the appropriateness of each part of the test which brings about the composite scores (Lee, 2006; Brennan, 2011). The important points indicating appropriateness of a test with multiple content domains or multiple related traits are: (a) estimation of the reliability of the composite score applying different weighting schemes, and (b) the universal score that shows if there is a true relationship between the scores obtained from a multitask speaking measurement (Brennan, 2001). In addition, it also determines reliability, specifies appropriate strategies for further improvement, and makes components of the assessment as well as the whole test more reliable (Burch, 2008). Therefore, the new method of measurement should help to increase the validity and reliability in English language measurement and evaluation, and make the assessment errors known (Gebril, 2010).

It can be concluded that data analysis related to English speaking has to be carried out by assessing examinees or students' tasks or practical works using reliable criteria. It also requires assessment by an expert. Applying multivariate generalizability theory to determine the validity of an English speaking skills test, in which analytic rating scale is used, is appropriate. The objectives of this research study were: (a) to determine the reliability coefficients of an analytic rating scale designed to assess the English speaking skills of Grade 6 Thai students, and (b) to compare the reliability coefficients of this assessment with different numbers of speaking tasks and raters. The present study addressed the following research questions:

1. How valid is the analytic rating scale?
2. How does the reliability coefficient change if the number of speaking tasks increases?
3. How does the reliability coefficient change if the number of raters increases?

METHODS

Participants

Participants consisted of 101 Grade 6 students (44 of whom were boys and 57 – girls) in a school of government under the Office of Basic Education Commission located in Bangkok. The average age was 12 years old. All of them studied English as a foreign language at school taught by Thai and foreign teachers.

Instrument

Analytic rating scale was used to score the English speaking ability in this research. The scale was specifically developed to assess English speaking skills of grade 6 Thai students who studied English as a compulsory subject and as a foreign language at school. To develop the English speaking skills analytic rating scale, the researcher studied the learning standards and the indicators of the foreign language (English) learning area of the national core curriculum for basic education (B.E.) 2551 (2008 A.D.), as well as the components of language ability presented by Bachman and Palmer (2010), and also interviewed nine English teaching experts. These experts included English teachers (primary level), experts in English speaking skills teaching, experts in English assessment, and experts in educational assessment. After that, the analytic rating scale was drafted and was verified by the experts before trying it out with the students who were in the samples.

After the rating scale was verified by the experts, it was piloted with 101 students. The English speaking skills of the students were tested through speaking tasks. The tasks were 3-minute oral presentations on two topics: Myself and My Favorite Person. Before the speaking test, the students were given a chance to ask questions. The students used a headset to record their voices in a computer. After the students completed the speaking test, three raters were given the sound records of all 101 participants to score the two speaking tasks (Myself and My Favorite Person) independently. The analytical scale consists of 6 components: vocabulary, syntax, cohesion, pronunciation, ideational function and fluency. The score 1–4 indicates very poor, poor, moderate or good performance, respectively.

Procedure

Raters. Three raters participated in this research. All of them were second- or third-year Ph. D. students in the English as an International Language Program in Chulalongkorn University. They were also English teaching lecturers at a tertiary education level. The researcher developed an assessment manual for the raters which included the objectives of the assessment, description of the tasks to be assessed, the rating procedures, and the rating criteria. The raters attended a rater training session to increase inter-rater reliability. During the training, individual raters listened to recorded voices of 6 students and then rated them on six assessing aspects. Then they discussed similarities and differencies in their scoring with explanations before the scores were agreed upon. After that, each of them was given the recorded voices of 10 students for individual rating before their rating scores were compared and discussed.

Data analysis. The data analysis was made using the computer program mGENOVA (Brennan, 2001). It was used to estimate the variance and covariance components and the reliability coefficients of the subsections scores and the composite score in the present study's analytic rating scale. The chosen G-study design for the present study is a two-facet crossed design with tasks and raters ($p^{\bullet} \times t^{\bullet} \times r^{\bullet}$).

In the design $p^{\bullet} \times t^{\bullet} \times r^{\bullet}$ it is assumed that all of the students (*p*) are tested in all tasks (*t*) and all the tasks are scored by the same rater (*t*).

RESULTS

Research Question 1: How valid is the analytic rating scale?

1.1 Estimated variance and covariance components (G-study). The variance and covariance components obtained from the multivariate generalizability analysis (G-study) of the speaking analytic rating scale, and the universe score correlations between subsections of the analytic rating scale that were estimated $p^{\bullet} \times t^{\bullet} \times r^{\bullet}$ by design are as shown in Table 1. It was found that the "person" variances have the greatest variance components among the subsections of the analytic rating scale. They accounted for 65.64%, 61.66%, 44.83%, 61.61%, 78.98% and 78.43% of the total variance in vocabulary, syntax, cohesion, pronunciation, ideational function and fluency respectively. The second were the variance components of "person by task by rater" interaction, followed by "task by rater" interaction, "person by task," and "person by rater" respectively. The results indicate that speaking proficiency depends on person**.** Both task and rater showed no major impact. As it can be seen, the variances of both "person by rater" and "person by task" interactions were lower than 1% in every component except for the vocabulary component of "person by task" which accounted for 4.57%.

As for the results related to the main effect of task and rater, it was found that the greatest variance of 35.24% was the components of task of cohesion. The second greatest was the syntax component which accounted for 8.95%. However, they were smaller than that of person. This shows that there was an error by task in the speaking ability of a person in syntax and cohesion. Concerning rater, it was found that every component was lower than 1% except for the syntax variance component which accounted for 4.09%. This shows that there was an error in the scoring practice of each rater regarding syntax.

The variances accounted for of (a) task by rater interaction, and (b) person by task by rater interaction, were found to be between .87% and 33.36%. These non-zero variance components indicated that variation in speaking proficiency of a person was the result of differences in the rank-ordering of task and rater, and/or errors from each rater and each task.

It was found that correlation between the components of speaking skills vary between medium and high levels (.747–.929). This demonstrates that speaking ability of the students depends on all the six components, not on any one in isolation. It was found that the ability on syntax highly correlated with every component and was related to fluency the most at .929. As for vocabulary, it was found to highly correlate to cohesion and ideational function (.880–.833) but it had medium correlation with syntax and pronunciation (.784 and .747 respectively). Concerning ideational function, it was found to highly correlate with vocabulary, syntax, cohesion and pronunciation (.833, .817, .822 and .860 respectively). This means that students who have high ability on ideational function must have high ability on vocabulary, syntax, cohesion and pronunciation. It was also found that cohesion has high correlation with vocabulary, syntax, ideational function (.880, .810 and .882 respectively), whereas fluency correlated to vocabulary, cohesion and ideational function at medium level (.755, .792 and .786 respectively) but highly correlated with syntax and pronunciation (.929 and .883).

Table 1. *Variance (in bold) and covariance components with correlations (in italics), for the person by task and by rater analysis*

Source of variation	Estimated variance component (in bold)											
	Voc.	%	Syn.	%	Coh.	%	Pro.	%	Fun.	%	Flu.	%
Person (p)	**.333**	65.64	.784		.880		.747		.833		.755	
	.202		**.199**	61.66	.810		.924		.817		.929	
	.312		.222		**.377**	44.83	.755		.822		.792	
	.160		.153		.172		**.138**	61.61	.860		.883	
	.300		.227		.315		.199		**.389**	78.98	.786	
	.300		.285		.335		.226		.337		**.474**	78.43
Task (t)	**.000**	.00										
	.025		**.029**	8.95								
	.053		.098		**.297**	35.24						
	.026		.017		.050		**.000**	.00				
	.032		.025		.095		.015		**.000**	.00		
	.025		.020		.095		.008		.008		**.000**	.00
Rater (r)	**.000**	.00										
	.008		**.013**	4.09								
	.044		.007		**.000**	.00						
	.025		.006		.018		**.000**	.00				
	.022		.004		.053		.018		**.000**	.00		
	.008		.004		.022		.000		.022		**.000**	.00
pxt	**.023**	4.57										
	.015		**.000**	.00								
	.006		.015		**.000**	.00						
	.010		.004		.010		**.000**	.00				
	.014		.001		.001		.011		**.000**	.00		
	.012		.032		.037		.002		.013		**.000**	.00
pxr	**.000**	.00										
	.003		**.000**	.00								
	.024		.020		**.000**	.00						
	.019		.025		.014		**.000**	.00				
	.011		.016		.019		.010		**.000**	.00		
	.024		.039		.033		.025		.024		**.000**	.00
txr	**.079**	15.46										
	.016		**.003**	.87								
	.023		.009		**.051**	6.11						
	.029		.006		.017		**.011**	5.02				
	.036		.010		.048		.019		**.046**	9.37		
	.019		.001		.027		.001		.018		**.027**	4.44

ptr,e	**.073**	14.33										
	.020		**.079**	24.42								
	.039		.034		**.116**	13.83						
	.034		.041		.021		**.075**	33.36				
	.026		.021		.031		.019		**.057**	11.65		
	.033		.052		.040		.036		.031		**.104**	17.13
Total	.508	100	.323	100	.841	100	.224	100	.492	100	.604	100

Notes: The values along the diagonal (show in bold) represent the variance components, while the lower of the diagonal is the covariance, and that on the top of the diagonal (shown in italics) is the correlation.
Voc. – vocabulary; Syn. – Syntax; Coh. – Cohesion; Pro. – Pronunciation; Fun. – Ideational function; Flu. – Fluency

1.2. Estimated composite score.

1.2.1 Reliability of the individual rating scales. The reliability of the speaking analytic rating scale was analyzed by multivariate generalizability theory yielding results as shown in Table 2**.** It was found that when a priori weight was applied, the highest universe score variance was fluency, which accounted for 20.82% of the total variance. This suggests that fluency has the greatest impact on variance of the composite score of speaking skill, followed by ideational function, cohesion, vocabulary, syntax and pronunciation, respectively. As for relative error variance, it was found that the highest error variance at 25.57% was vocabulary, followed by pronunciation, cohesion, ideational function, syntax and fluency respectively. Concerning absolute error variance, it was found that the highest error variance at 42.62% was cohesion, followed by vocabulary, ideational function, pronunciation, syntax and fluency respectively. It was found that the value of Gen Coefficient of every component of speaking analytic rating scale in relation to reliability coefficient was from medium to high levels (.649–.872) with the highest reliability coefficient being ideational function followed by fluency with a close value (.872 and .821 respectively). The values for vocabulary, cohesion and syntax were close (.776, .764 and .716 respectively). The lowest value was for pronunciation (.649). It was also found that the Phi Coefficient was at medium level for every component (.656, .617, .616, .790, and .784) in vocabulary, syntax, pronunciation, ideational function and fluency, respectively, except for cohesion which was at the low level (.448).

1.2.2 Reliability of the composite score. In relation to reliability of the composite score of speaking analytic rating scale, the relative error variance and absolute error variance were found to be .022 and .034 respectively. As for the reliability coefficient, both the generalizability coefficient and Phi Coefficient were high at .924 and .885 respectively as shown in Table 2.

Table 2. *Composite score analysis result*

	English Speaking Skill components					
Contributions to	Voc.	Syn.	Coh.	Pro.	Fun.	Flu.
A priori weight	.166	.166	.166	.166	.166	.166
Universe score variance (%)	17.10	13.71	18.44	11.14	18.79	20.82
Relative error variance (%)	25.57	14.12	17.55	20.08	15.42	7.26
Absolute error variance (%)	22.78	2.21	42.62	13.08	18.63	.67
Gen Coefficient	.776	.716	.764	.649	.872	.821
Phi Coefficient	.656	.617	.448	.616	.790	.784

Composite	
Relative Error Variance	.022
Absolute Error Variance	.034
Generalizability coefficient	.924
Phi	.885

Notes: Voc. – vocabulary; Syn. – Syntax; Coh. – Cohesion; Pro. – Pronunciation; Fun. – Ideational function; Flu. – Fluency

Research Question 2: How does the reliability coefficient change if the number of speaking tasks increases (D-study)?

In order to answer the question of how to design a speaking skill test applying an analytic rating scale to get a dependable reliability coefficient by using various forms of $p^{\bullet} \times t^{\bullet} \times r^{\bullet}$ design, the researcher compared the generalizability coefficient and Phi coefficient in different number of tasks based on three raters. The findings were that for the generalizability coefficient, five or more tasks should be used for all of the six test components to be reliable (reliability coefficients = 0.80 and above) as shown in Figure 2. However, if a dependability of Phi Coefficient is required, six or more tasks can be used for every component, except for cohesion which should have 10 or more tasks. It can be concluded that in order to design a speaking skill test applying an analytic rating scale to get a dependable reliability coefficients, a minimum of 10 tasks is required as shown in Figure 3.

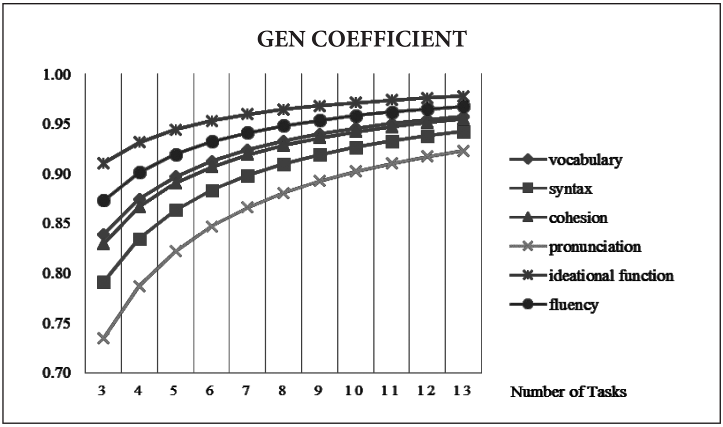

Figure 2. *Generalizability coefficient for number of tasks from 3–13 (The number of raters is three).*

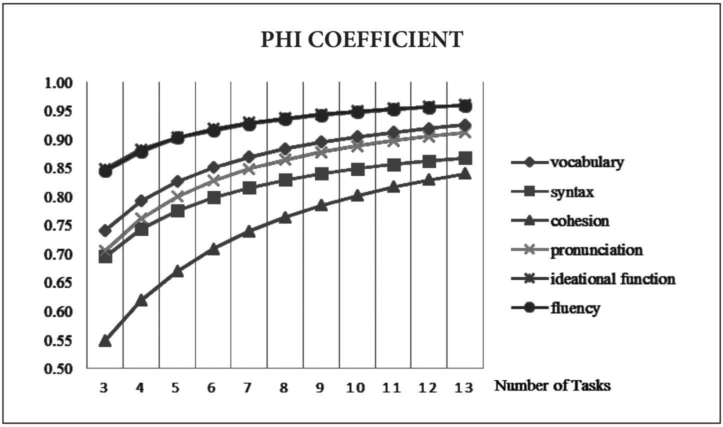

Figure 3. *Phi coefficient for number of tasks from 3–13 (The number of raters is three).*

Research Question 3: How does the reliability coefficient change if the number of raters increases (D-study)?

In order to answer the question related to the appropriate number of raters for a speaking skill test applying an analytic rating scale to get a dependable reliability coefficient by using various forms of $p^{\bullet} \times t^{\bullet} \times r^{\bullet}$

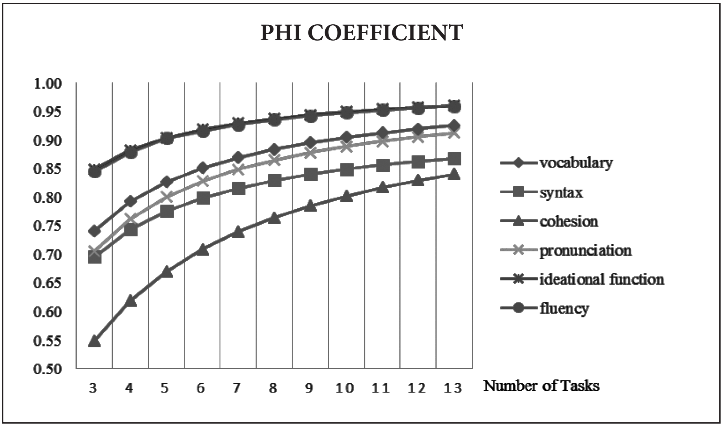

Figure 4. *Generalizability coefficient for number of raters from 3–13 (The number of tasks is three).*

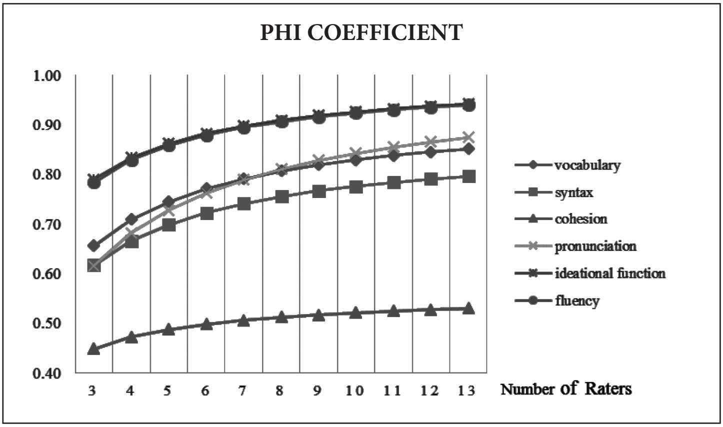

Figure 5. *Phi coefficient for number of raters from 3–13 (The number of tasks is three).*

design, the researcher compared the generalizability coefficient and Phi coefficient in different number of raters based on three tasks. The findings were that for the generalizability coefficient, six or more raters should be used to obtain high reliability coefficients (reliability coefficients = .80 and above). However, if a dependability of Phi Coefficient is required, eight or more raters should be used for every component, except for cohesion which was found to have low reliability coefficients even when more than 10 raters were used as shown in Figure 4 and Figure 5.

Therefore, it can be concluded that to design a speaking skill test applying an analytic rating scale to get dependable reliability coefficients, a relatively large number of tasks (minimum of six tasks) is required, whereas the number of raters can be as small as three.

DISCUSSION

The objectives of this study were to determine the reliability coefficients of an analytic rating scale designed to assess English speaking skills of Grade 6 Thai students, and to study the impact on reliability coefficients when the numbers of speaking tasks and raters are different. The research findings can be discussed in detail as follows:

1. The greatest variance component of the composite score of the analytic rating scale was fluency followed by ideational function, cohesion, vocabulary, syntax, and pronunciation respectively. The findings of this study are consistent with the previous research studies that found fluency to be the most important component of oral proficiency (Sato, 2011; Iwashita & Grove, 2003). In this study, it was found that fluency has the greatest variance component of the composite score of the analytic rating scale which indicated that fluency was the most important component of the rating scale. This means that good English speaking or communicative speaking does not only depend on the accuracy of grammar or vocabulary, but also on fluency. Therefore, teachers should see the importance of fluency as well. The second most important components are ideational function and cohesion, demonstrating that it should be also important to know what the student wants to communicate (Iwashita & Grove, 2003). Children must be able to make others understand what they want, what they think, and convey what is the real aim of each communication by learning from their past experiences. Children must understand the use of conjunctions which are required to make others understand the connection of ideas. They should also know and be able to choose appropriate vocabulary and use it in various situations. Concerning syntax and pronunciation, the variance was found

to be the smallest, indicating that when assessing the students' spoken English at this age, the raters saw the importance of fluency, ideational function, and cohesion more than vocabulary, syntax and pronunciation. This is in line with the previous research reported by Sato (2011), and de Jong and van Ginkel (1992), which paid more attention to fluency and content than vocabulary, grammar and pronunciation.

2. The composite score reliability of the speaking analytic rating scale was found to be at a high level (above .80). This shows that the analytic rating scale used for the assessment of grade 6 students' English speaking skills is appropriate and reliable. This might be due to the use of 4-point rating (scores in each component of the speaking skill ranging from 1–4 indicating very poor, poor, moderate and good, respectively). Luoma (2004) stated that a speaking analytic rating scale should consist of 4–6 levels of measurement for the convenience of raters. It is also in line with Weigle (2002) who stated that most analytic rating scales were not employing more than 4 levels of measurement which made it easier for raters to rate. A clear definition was given in each level which helped all raters agree on the rating framework. Past research studies found that "speaking analytic rating scales" usually have 4–5 levels of measurement, and they usually consist of 4–6 components of speaking skill (Lee, 2006; Sato, 2011; Sawaki, 2007).

3. The size of error variance of tasks in this study was greater than that of the raters in the cohesion and syntax components (Table 1). It might be due to the varying levels of difficulty of the tasks, which impacted students' speech production. Students might not be able to make a linkage in their speaking and could not construct a good grammatically correct sentence for the more difficult tasks or the tasks related to an unfamiliar situation. This is in line with the previous research which found that tasks had more error variance than raters (Lee, 2006; Sawaki, 2007).

4. Increasing the number of tasks and raters increased the generalizability coefficient and Phi coefficient**.** However, a minimum of six tasks and 6–10 raters were needed to obtain satisfactory reliability coefficients for each component. Increasing the number of tasks increased the score dependability more than increasing the number of raters. This is in line with the previous research which found that increasing the numbers of tasks and raters would increase the score dependability. The finding of this study is also in line with Gebril's (2010) study on English writing

scoring schemes which found that increasing the numbers of tasks and raters increased the generalizability coefficient and Phi coefficient. However, increasing the number of raters in practice, would increase the cost of assessment. This study revealed that generalizability coefficient and Phi coefficient can be increased to a reliable level by increasing the number of tasks, which would be more cost-effective than hiring and training additional raters.

APPENDIX

English speaking tasks

The task used in the assessment of English speaking of grade 6 students is in the form of oral presentation. It consists of 2 topics: Myself and My favorite person.

Assessment criteria

An analytical rating scale was used. The scopes of assessment are:

1. Vocabulary: Concerning vocabulary aspect, the scoring determines 2 things out of the students' speaking, the range and the accuracy of the vocabulary used. Students use a variety of vocabulary and have the ability to use accurate and appropriate vocabulary to convey the message they want to communicate.

2. Syntax: The score is given for the construction of sentences. It is determined from the ability to construct a grammatically correct sentence, such as appropriate using of a subject, a verb and an object, with the correct meaning they want to communicate.

3. Cohesion: The cohesion is determined from the ability to link the speaking elements into a story with appropriate relationships.

4. Pronunciation: The scoring in pronunciation aspect is determined from the students' ability to correctly express the stress, sound segmentation, and intonation that makes their speaking clear and understandable by the audiences.

5. Ideational function: Concerning ideational function aspect, the scoring is determined from the students' ability to correctly convey desirable messages. The speaking is understandable by the audiences. For example, students keep speaking relevant to the title, answer the questions and give appropriate answers to the questions.

6. Fluency: Fluency in English speaking can be determined from students' ability to express fluent and smooth speech in English. They do not show stumbled speech, irregularity or break within the flow of speaking.

The scoring system 1–4 indicates very poor, poor, moderate and good performance respectively. The details are shown below:

Analytical rating scale	Level of ability	Descriptions
1. Vocabulary	Very poor (1)	– Very limited ability in using vocabulary – Used repeated vocabulary in every or almost every sentence – Used incorrect or inappropriate vocabulary
	Poor (2)	– Limited ability in using vocabulary, no variety of vocabulary used – Often used repeated vocabulary – Often used incorrect or inappropriate vocabulary
	Moderate (3)	– Used variety of vocabulary – Sometimes used repeated vocabulary – Used correct and appropriate vocabulary, but there were times when incorrect or inappropriate vocabulary was used
	Good (4)	– Used variety of vocabulary – Used correct and appropriate vocabulary most of the time. Very few mistakes or no mistakes were made.
2. Syntax	Very poor (1)	– Spoke in single words, could not put words in a sentence – Used incorrect structure at all time
	Poor (2)	– Spoke in sentences – Often used incorrect structure that could not make the audiences understand clearly
	Moderate (3)	– Spoke in sentences – Used incorrect structure sometimes but understandable
	Good (4)	– Spoke in sentences – Used correct structure most of the time. Very few mistakes were made

3. Cohesion	Very poor (1)	– Used very few conjunctions or no conjunction at all – Almost all of the conjunctions used were incorre ct
	Poor (2)	– Few conjunctions were used – Conjunction usage was often confusing and caused misunderstanding
	Moderate (3)	– Correct conjunctions were used – The conjunction used might cause a little confusion in sentences but no misunderstanding
	Good (4)	– Correct conjunctions were used – Correct and clear most of the time, no misunderstanding, very few mistakes made or no mistakes at all
4. Pronunciation	Very poor (1)	– Nearly all pronunciations were incorrect and not understandable – Spoke very little or not at all
	Poor (2)	– Incorrect pronunciations most of the time – A lot of mistakes made in stress, sound segmentation and intonation which caused misunderstanding or the message was not understandable
	Moderate (3)	– Incorrect pronunciations sometimes – Some mistakes made in stress, sound segmentation and intonation but the message was understandable
	Good (4)	– Most pronunciations were correct or no mistake made – The message wished to convey were easily understood by the audiences
5. Ideational function	Very poor (1)	– Off-topic or not spoken at all – Nearly all of the message was not understandable
	Poor (2)	– The content was hardly relevant to the topic – The objective of the speaking was poorly understood
	Moderate (3)	– The content was relevant and appropriate although – some mistakes were made – The objective of the speaking was understood
	Good (4)	– The content was relevant and appropriate, very few mistakes were made or no mistake made – The objective of the speaking was easily understood

6. Fluency	Very poor (1)	– Stammered in single words – No sentence was spoken out at all
	Poor (2)	– Very often speech was slow and not smooth – Unusual pauses made, unfinished sentence made which causes missing or incomplete content
	Moderate (3)	– Pauses made, not smooth speech sometimes – Got stuck or pause to think sometimes – Finished the sentences with complete content
	Good (4)	– Constant speaking speed, natural and smooth, no stammering – Spoke without getting stuck – Finished the sentences with complete content

References

Bachman, L. F., Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: OUP.

Barton, J., Collins, A. (1997). *Portfolio Assessment: A handbook for Education*. California. Addison-Wesley Publishing Company.

Brennan, R .L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education*, 24 (1), 1–21.

Cooper, P. L. (1984). *The assessment of writing ability: A review of research (GRE Board Research Report No. GREB 82-15R/ETS Research Report No. 84-12)*. Princeton, NJ: Educational Testing Service.

de Jong, J. H. A. L., van Ginkel, L. W. (1992). Dimensions in oral foreign language proficiency. In L. Verhoeven & J. H. A. L. de Jong (Eds.), *The construct of language proficiency: Application of psychological models to language assessment*, 187–206. Amsterdam: John Benjamins.

Efthymiou, G. (2012). Portfolio Assessment of Speaking Skills in English as a Foreign Language in Primary Education. *Research Papers in Language Teaching and Learning*, 3 (1), 200–224.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing*, 26, 507–531.

Gebril, A. (2010). Bringing reading-to-write and writing-only assessment tasks together: A generalizability analysis. *Assessing Writing*, 15, 100–117.

Huot, B. A. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, 60, 237–263.

Iwashita, N., Grove, E. (2003). A comparison of analytic and holistic scales in the context of a specific-purpose speaking test. *Prospect*, 18 (3), 25–35.

Lee, Y .W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing*, 23 (2), 131–166.

Lee. Y.-W., Kantor, R. 2005: *Dependability of new ESL writing test scores: Evaluating prototype tasks and alternative rating schemes*. (TOEFL Monograph Series, MS-31). Princeton, NJ: Educational Testing Service.

Lisa, A. K., Brian, E. C., & David B. S. (2010). Using multivariate generalizability theory to assess the effect of content stratification on the reliability of a performance assessment. *Adv in Health Sci Educ*, 15, 717–733.

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge university press.

Sato, T. (2011). The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing*, 29 (2), 223–241.

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing*, 24 (3), 355–390.

Webb, N., Shavelson, R., & Haertel, E. (2007). Reliability coefficient and generalizability theory, *Handbooks of Statistics 26* [Online]**.** Retrieved from http://www.stanford.edu/dept/SUSE/SEAL/Reports_Papers/methods_papers/G%20Theory%20Hdbk%20of%20Statistics.pdf.

Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

Westervelf, F. M., Moran, A. C. (2011). Expository language skills of young school-age children. *Language, Speech, and Hearing Services in Schools*, *42,* 182–193.

ACKNOWLEDGEMENTS

The researchers would like to thank THE 90th ANNIVERSARY OF CHULALONGKORN UNIVERSITY FUND (Ratchadaphiseksomphot Endowment Fund) for funding this research.

ŠEŠTOKŲ TAILANDIEČIŲ ANGLŲ KALBOS GEBĖJIMŲ ĮVERTINIMAS: DAUGIAMATĖS GENERALIZACIJOS TEORIJOS TAIKYMAS

Daruwan Srikaew, Kamonwan Tangdhanakanond, Sirichai Kanjanawasee
Chulalongkorn universitetas, Tailandas

Santrauka. ***Įvadas, tikslas.*** Šiame tyrime nagrinėjama analitinė rangavimo skalė anglų kalbos kalbėjimui vertinti, sukurta šeštokams tailandiečiams, besimokantiems anglų kaip antrosios kalbos. ***Metodai.*** Šiam tyrimui buvo sukurta ir validizuota vertinimo skalė, taip pat atliktas žvalgomasis vertinimas naudojant 101 mokinio, lankančio valstybinę mokyklą Bankonke, Tailande, duomenys. Sukurta analitinė vertinimo skalė vertina 6 komponentus: žodyną, sintaksę, rišlumą, tarimą, turinį ir sklandumą. Vertinimo skalės patikimumas analizuotas naudojant skirtingą kalbėjimo užduočių kiekį ir skirtingą vertinančių ekspertų skaičių. Analizuojant rezultatus buvo remiamasi Daugiamate generalizavimo teorija (G teorija). ***Rezultatai.*** Rezultatai rodo, kad sklandumas sudaro didžiausią analitinės vertinimo skalės sklaidos dalį, toliau pagal paaiškinamos sklaidos dalį seka turinys, rišlumas, žodynas, sintaksė ir tarimas. Bendros skalės patikimumas buvo aukštas (daugiau kaip 0,80). Nustatyta, kad patikimumo koeficientai kiekvienam komponentui yra aukšti (daugiau kaip 0,80), kai vertinamos šešios ar daugiau užduočių ir tai atlieka 6–10 ar daugiau vertintojų. Sąsajos didėjo, kai buvo didinamas vertinamų užduočių arba vertintojų skaičius. Taip pat išsiaiškinta, kad pakankamai aukštas patikimumo Phi koeficientas (daugiau kaip 0,80) pasiekiamas naudojant tik 6 užduotis ir 3 vertintojus. ***Išvados.*** Pagrindiniai rezultatai aptariami detaliau.

Pagrindiniai žodžiai: kalbėjimo angliškai įgūdžiai, analitinė vertinimo skalė, daugiamatė generalizacijos teorija.

Received: July 30, 2014
Accepted: January 26, 2015