Effects of Quantity and Quality of Speech on Group Oral Tests

Siwon Park
Yasushi Sekiya
Masaki Kobayashi
Yasuko Ito

Introduction

The primary purpose of the current study is to examine the extent to which raters' scores were affected by the quantity and quality of examinees' foreign language speech in group oral tests. Prior studies have suggested that the group oral test be a reliable testing technique. Those studies, however, mostly concerned test validation using rating scores without fully addressing how the quantity and quality of the speech produced by examinees may affect raters' judgments. Researchers, such as Hildon (1991) and Fulcher (1996), were active advocates of the group oral tests. However, a few recent validation studies (Kobayashi, Johnson, & Van Moere, 2005; Nakatsuhara, 2010; Park, 2008; Van Moere, 2006; Van Moere, 2010) have expressed reservations about the use of the test especially for high-stakes testing.

Among the researchers who have explored group oral tests, Hildon (1991) appears to be the first who formally mentioned the use of the group oral test in oral assessment. Hildon points to several advantages of the test while justifying its use in Zambia as part of school exams. He argues that group tests are economical relative to conventional interviews since large numbers of candidates can be heard in a short time. Also, the test would suggest several advantages for testing children's oral ability, especially for the

shyer or more nervous ones. In his trial of the group oral exam in Zambia, however, Hildon noticed a couple of problems in administrating and scoring the exam which included the issue of content and questions of cultural appropriateness in addition to the reliability in rating and standardization of the task itself.

Kobayashi, Johnson, and Van Moere (2005) studied the relationship between the amount of students' output amounts and their scores in group oral tests administered yearly at a university in Japan. Their study, similar to the current one in its purpose, examined if the amount of speech defined by the number of words spoken and its quality defined by vocabulary breadth are associated with the scores assigned by the raters. They found that there was a systematic relationship between the amount of speech and the scores: the more the learners spoke, the higher scores they received. Van Moere (2006) took a more extensive look at the validity of group oral tests. He conducted a G-study to locate the sources of variation in test scores and found that person-by-occasion was the greatest source of variance, while topic was not a significant factor. Van Moere argues that his G-study revealed that the variations in performances themselves were more responsible for the differences in test scores from one occasion to the other.

Nakatsuhara's (2010) study on group oral tests concerns more practical aspects of the test and provides more pertinent suggestions to the administration of the tests. She argues that in order to control the extroversion levels of examinees, a test group must involve no more than three examinees. She notes in her study that the number of participants in a group oral test significantly affects the group dynamic. When four participants sat the exam, the discussion turned into a presentation event, in which each participant, without exchanging turns, presented his/her opinion and passed the turn to the next participant. In addition to limiting the number of participants, Nakatsuhara recommends using more closed, goal-oriented tasks in a group oral test,

such as information gap or picture difference tasks. This is to force all participants to attend to the oral performance equally contributing to the completion of the task(s). Such use of more goal-oriented tasks in group oral tests was strongly advocated also by Van Moere (2010) and Park (2008) as the tasks facilitate more negotiation of meaning among participants, which is closer to authentic conversations.

Concerned with the increasing popularity of group oral tests in language education, more validation studies on the tests are called for. The current study aims to add a piece of validity evidence to prior studies for the use of the tests. For such a research purpose, the following research questions were to be addressed in the study:

1. Does the amount of speech have a considerable influence on group oral rating?
2. Does the linguistic quality of speech meaningfully influence group oral rating? If so, what aspect(s) is particularly influential – accuracy, complexity, and/or lexical diversity?
3. Which aspect of speech influences raters' judgments more in group rating – linguistic quality or quantity?

By addressing the three research questions, we will be able to examine the extent to which the linguistic quantity and quality of L2 examinees' English speech affect raters' score assignment in group oral tests.

Methodology

1. Participants and speech sample data

The speech samples used for the current study come from 11 group oral tests of an in-house English proficiency test (known as Kanda English Proficiency Test; KEPT)[1]

[1] Interested readers in the in-house proficiency test may refer to Bonk and Ockey (2003) for the details of the test.

administered at a Japanese university in 2008. Of the 11 tests, seven included four students and four included three students. Thirty students were female, and the rest male (Female=30; Male=10). Among the 40 students, 12 were first year, 22 second year, and the rest third year students (freshmen=12; sophomores=22; juniors=6).

2. Procedures

More than a hundred group oral tests were video-recorded at the 2008 test administration, and 11 of them were randomly selected and transcribed for subsequent coding. For the coding of the numbers of words and turns, the coding scheme was adopted from Kobayashi et al. (2005), which they developed and used in their study. For instance, "a turn was defined as consisting of the time from when the speaker first begins an utterance until the time another speaker replies, comments or interrupts" (p.279). Also, only complete words were counted as a turn or word. Interjections, simple back-channeling, or repetitions were not counted as turns, i.e., only meaningful utterances were counted as turns.

For the coding of linguistic measures, the primary coder read and coded all the speech samples, while the second coder coded only 20% of the speech data. Upon the completion of all linguistic coding, the inter-coder reliability was checked between the first and second coders. Any discrepancy between the two coders was resolved through discussion based on the coding guidelines that they were asked to utilize (See Appendix 1 for the actual coding guidelines).

Together with the quantity and quality indexes identified through coding procedures, test scores were entered into the analysis that were assigned by two raters and statistically adjusted for their fairness. All the measures and scores were subsequently analyzed using the *vocd* (McKee, Malven, and Richards, 2000) modeled under the CLAN program (McWhinny, 2000), EXCEL, and SPSS.

Analysis

Two types of measures, quantity and quality, and rating scores were entered into the analysis. For the quantity estimation, numbers of words and turns were calculated and entered into the analysis. As the quality measure, the following six units of analyses and measures were identified and calculated:

- Number of T-units
- Number of clauses
- Mean number of clauses per T-unit (complexity)
- Percentage of error-free T-units (accuracy)
- *D* (lexical diversity)
- TTR (lexical diversity)

T-units and clauses were defined following the coding guidelines in Appendix 1, and as an index of lexical diversity, *D* values were calculated using the *vocd* (McKee, Malven, and Richards, 2000) modeled under the CLAN program (McWhinny, 2000). In the actual analysis, only the linguistic measures further explained in Table 1 were applied.

Table 1 Measures entered into the analyses

Category	Feature	Unit of Analysis
Grammar		
Accuracy	Global accuracy	Percentage of error-free T-Units
Complexity	T-Unit complexity ratio	Mean number of clauses per T-Unit
Lexical diversity	Mathematical modeling of how new words are introduced into larger and larger language samples	*D* values

The oral rating scores used in the analyses were all double-scored and Rasch-adjusted for rater severity. The rating was done using an analytic scale of five proficiency categories (Pronunciation, Fluency, Grammar, Vocabulary, and Communicative effectiveness). For our research purpose, we decided to prepare two sets of total scores – total scores of all five categories (as Total-5) and total scores of only three categories (fluency, grammar, and vocabulary; as Total-3). The total scores of the three categories were prepared considering the comparability of the scores and the linguistic measures that were entered into the correlational analyses including the regression analyses.

Results

With the arranged data, we ran a series of correlational analyses. First, bivariate correlations across different measures and score variables were examined to check if the variables were systematically related to each other. Next, a series of multiple regressions followed, and the outputs were examined to determine the extent to which the independent measurement variables predict the total score variables.

1. Bivariate correlations

Table 2 presents the first result of the correlations between the three category scores of the group oral test and the three linguistic measures.

Table 2 Bivariate correlations

	Group oral		
	Vocabulary	*Grammar*	*Fluency*
D	.294	.288	.391*
Accuracy	.311	.352*	.247
Complexity	-.168	-.084	-.101

* $p < .05$

Out of the nine comparisons, two correlations were found significant between Accuracy and Grammar, and *D* values and Fluency, while *D* values did not correlate with Vocabulary. In addition, Complexity did not correlate with Grammar; the coefficient is close to zero, indicating that essentially there is no relationship between the two variables. Furthermore, even the significant correlations were marginal in their size. Table 3 reports the result of the second correlational analysis.

Table 3 Correlations across all the measurement variables

		1	2	3	4	5	6	7
1	*Oral total (5)*	1.00						
2	*Oral total (3)*	.98*	1.00					
3	*Accuracy*	**.31***	**.33***	1.00				
4	*Complexity*	-.13	-.13	**-.32***	1.00			
5	*D*	**.35***	**.36***	-.02	.09	1.00		
6	*# of turns*	.30	.28	.28	-.20	**.44***	1.00	
7	*# of words*	**.48***	**.47***	.21	-.04	**.49***	.81*	1.00

* $p < .05$

The correlation coefficients presented in Table 3 show the relations across all the measurement variables including the two total scores of the group tests. The values in bold are some of the significant correlation coefficients directly related to one of our research questions.

Among the linguistic variables, first, the accuracy measure is correlated significantly with the two totals, while Complexity does not. Interestingly, the complexity measure is correlated negatively with the accuracy measure. Also, the coefficients between the lexical measure, *D* and the total scores are significant, so as the ones between the accuracy and the total scores.

Among the quantity measures, only the number of words is correlated significantly with the total scores. Also, the number of words and the number of turns are correlated significantly with the *D* values. Considering that *D* is a measure of lexical density, the significant and medium size correlation between *D* and the number of words and the number turns is reasonable. Finally, among the quantity and quality measures, the number of words resulted with the largest correlations with the two total scores.

2. Multiple Regressions

Table 4 reports the result of the first regression analysis. Total-5 was entered as the dependent variables and two amount variables and three linguistic variables as predictors into the equations. The regression analysis was conducted in the partly sequential manner, i.e., by adding additional predictor variables one at a time; the effect was examined for information in addition to the variable entered earlier.

Before the variables were entered into the regression analyses, each variable was checked for their normality of the data. As the distributions of the number of words and the number of turns were not normal, the two variables were transformed to correct their non-normality using Square-root and Log transformation. The variable names, SRwords and LOGturns in Table 4 specify the transformed nature of the variables. In addition, unstandardized regression coefficients, standardized coefficients (beta), and semi-partial correlations are each presented. Since the first research question of the current study is concerned with the effect of the amount of speech in group oral tests, the number of words was entered into the regression equation first.

Table 4 Step-wise regression with quantity and quality variables

Model		*B*	*SE B*	*Beta*	Part
1	(Constant)	10.127	1.032		
	SRWORDS	0.280	0.083	0.479*	0.479
2	(Constant)	10.346	1.122		
	SRWORDS	0.322	0.116	0.551*	0.400
	LOGTURNS	-0.702	1.331	-0.105	-0.076
3	(Constant)	10.575	2.420		
	SRWORDS	0.327	0.131	0.560*	0.345
	LOGTURNS	-2.200	1.570	-0.328	-0.194
	Accuracy	0.028	0.016	0.263	0.243
	Complexity	-1.168	1.096	-0.184	-0.147
	D	0.030	0.026	0.185	0.159

Note *DV=Total*-5. $R^2 = .230$ for Step 1; $\Delta R^2 = .006$ for Step 2; $\Delta R^3 = .113$ for Step 3. * $p < .05$

The R-squared of Model 1 (only with the number of words variable) shows that more than one fourth of the variability in Total-5 is predicated by the number of words alone. Entering additional variables into SRWORD does not help account for the variability of Total-5. Adding the number of turns to Model 1 does not nullify the significance of the number of words in Model 2. Such significance of the number of words remains even after adding other variables to the equation of Model 3. This finding suggests that the number of words is the single most significant variable that influenced the total scores in the group tests.

Another regression analysis was performed only with linguistic variables as the independent variables and the rating scores as the dependent variable. Table 5 reports the result. Subsequently, the number of words was entered into the equation step-wise, as in Model 2, so that the significance of the variable could be examined against other linguistic variables.

Table 5 Regression with linguistic variables and # of words

Model		*B*	*SE B*	*Beta*	Part
1	(Constant)	5.430	1.275		
	Accuracy	0.020	0.010	0.317*	0.300
	Complexity	-0.211	0.586	-0.055	-0.052
	D	0.035	0.014	0.366*	0.364
2	(Constant)	4.966	1.247		
	Accuracy	0.017	0.009	0.261	0.243
	Complexity	-0.167	0.564	-0.044	-0.041
	D	0.020	0.016	0.203	0.174
	SRWORDS	*0.115*	*0.057*	*0.328*	*0.279*

Note *DV = Oral total (5).* R^2 = .241 for Step 1; ΔR^2 = .078 for Step 2. * $p < .05$

As shown in Model 1, *D* (lexical diversity) was found to have influenced the rating most significantly, followed by Accuracy. Complexity resulted with nearly no impact on the rating, although it was to be reflected in Grammar and total scores. Moreover, adding the number of words to Model 1 nullifies the significance of Accuracy and *D* in Model 2, which confirms the finding of the first regression analysis, i.e., the strong impact of the number of words onto the rating scores.

Discussion and Conclusion

The main purpose of the current study was to examine the extent to which the linguistic quantity and quality of L2 examinees' English speech affect raters' score assignment in group oral tests. In order to achieve the purpose, speech samples were analyzed together with the rating scores. The analyses revealed important facets of group oral tests and suggest reconsiderations as to the use of the group oral test in L2 assessment.

The first research question asked if the amount of speech has a considerable impact on group oral rating, and the regression analyses revealed that to be the case. That is, the amount of speech was the single most important predictor of the rating scores. Furthermore, the effect of the amount of speech was not weakened even after other predictors were entered into the equation. Such a finding indicates that the linguistic quality of the examinees' speech may not be often appreciated by the raters. Additionally or alternatively, the descriptors of the rating scale may not have served raters to identify target linguistic features to evaluate. This masking effect of the speech amount is a serious concern as it defies the purpose of using the analytic rating scale that includes multiple traits of English proficiency.

In this study, we were also interested in whether or not the linguistic quality of speech meaningfully influences group oral rating. We were also concerned if the influence is to be statistically meaningful, what aspect(s) is particularly so – accuracy, complexity, and/or lexical diversity? The analyses revealed that Accuracy and the lexical density, *D* are correlated significantly with the total scores while Complexity is not. Complexity may be a difficult linguistic trait to assess in L2 learners' speech especially when raters are forced to evaluate the multiple aspects of the speech at once.

Finally, the study examined which aspect of speech – quantity or linguistic quality – influences raters' judgments more in group oral rating. The findings revealed that the amount of speech measured by the number of words had a significant impact on the ratings. This impact was larger than that of any other linguistic traits of the examinee speech measured in terms of accuracy, complexity, and lexical diversity.

In sum, the finding that the quantity meant much more to the raters than the quality of speech did not greatly surprise us. However, the size of its influence requires much closer attention to be paid to the practice of the test for educational purposes.

Continued effort for more rater training and continuous revision of the rating scale is a must in any testing practice. Together with such essential practices to improve the general aspects of oral testing, more specific considerations need to be given to the use of the group oral as an assessment technique. For instance, in group oral tests, equal participation in terms of the amount of speech must be encouraged for the examinees. It could be done through selecting appropriate test tasks and/or learner training before they sit to perform group discussion.

References

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

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

Hilsdon, J. (1991). The group oral exam: advantages and limitations. In J.C. Alderson and B. North (Eds.), *Language testing in the 1990s* (pp. 189-197). London: Modern English Publications and the British Council.

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, 16,* 275-295.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk (3rd Edition).* Mahwah, NJ: Lawrence Erlbaum Associates.

Nakatsuhara, F. (2010). *Interactional competence measured in group oral tests: how do test-taker characteristics, task types and group sizes affect co-constructed discourse in groups?* Paper presented at the Language Testing.

Ortega, L., Iwashita, N., Rabie, S., & Norris, J. M. (in preparation). A multilanguage *comparison of measures of syntactic complexity.* Honolulu: University of

Hawai'i, Foreign Language Resource Center.

Park, S. (2008). *An exploration of examinee abilities, rater performance, and task differences using diverse analytic techniques*. Unpublished doctoral dissertation, University of Hawaii, Honolulu, HI.

Van Moere, A., & Kobayashi M. (2003). *Who speaks most in this group? Does that matter?* Presented at the 2003 LTRC: Reading University.

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing, 23,* 411-440.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring Vocabulary Diversity Using Dedicated Software. *Literary and Linguistic Computing, 15,* 323-337.

Appendix 1

CODING GUIDELINES

The following definitions/guidelines are developed based on Brown, Iwashita, and McNamara (2005), Hunt (1970), Ortega, Iwashita, Rabie, and Norris (1998) and Sotillo (2000)

T-Units

- A T-Unit is defined: (1) an independent clause and all its dependent clauses and (2) as an independent clause only.

Examples:

(1) **(1 T-unit, 2 clauses)**

[I, I want to live in country in my future because hmmm... city is very noisy.]

(2-1) **(1 T-unit, 1 clause)**

[Ahh… I'm living in the city now.]

(2-2) **(2 T-unit, 2 clauses)**

[This morning at 7:30 he got up] [and... uh, he went to his office to work.]

- Do not count sentences fragments or incomplete sentence repetitions.

  Example:

  We can go to, [we can get anything we want.]

  and he happy, [he's happy to spend this year.]

- If a NP is standing alone or a subordinate clause is standing alone, do not count them as T-Units.

  Example:

  [I think... country, count... living in the country, person living in the country is more warm I think.]

  like place.] For example Disneyland or Disneysea,…

  Because the lady have a right to work.

- When there is a grammatical subject deletion in a coordinate clause, count the entire sentence as one T-Unit.

  Example:

  [ahh, he needs some money, and ... want to... his life more happier.]

- Count the following as subordinators: after, although, because, if, until, where, since, when, while, as if, as though, so that, in order that, so as, in order, as (many) as, more than, although, even though, despite, so (that).

Example:

[So... when I am old people, I want to live in...um...Nibu, my hometown.]

- Mark response formulas as separate T-units, so that they can be counted separately.

  Example: [Yeah], [Right]. [Thank you], [Yes, please], [Okay], [Sure], etc.

- Include incomplete starts in the same T-unit with following reformulations.

  Example:

  […living country] [ah no..., air of the country is so refresh, I think.]

Clauses

- **A clause** is a unified predicate containing a finite verb, a predicate adjective, or a nontarget-like predication in which the verb or part of the verb phrase is missing (Berman & Slobin, 1994).

- **A dependent clause** is a unified predicate (i.e., containing a finite verb, a predicate adjective, or a nontarget-like predication in which the verb or part of the verb phrase is missing) embedded in or dependent on a main matrix clause.

- **Finite clause:** A clause equals an overt subject and a conjugated verb, or a verb that is preceded by a modal (will, would, can, may, should, and so on).

  Example: "Japanese high school girls make a lot of money and buy Chanel, Gucci, etc." "I will visit my family next year."

  - Normally, a finite embedded clause is introduced by the complementizer that. (e.g., "I thought that things were not so difficult in this country"), while a finite

subordinate clause can be introduced by any of the above subordinators (see #5 in T-Units). (e.g., "When one comes to this country,…")

- Finite clauses can stand on their own as grammatical sentences or as the main clause of a larger clause if the complementizer is omitted. (e.g., "I studied medicine in my country.")

- **Nonfinite Clauses:** These types of clauses differ from the others in that they do not have an overt subject and the verb is preceded by to. Nonfinite embedded clauses are introduced by for, and although this complementizer is omitted sometimes, they cannot stand on their own as do finite clauses. ("We are going to live in San Francisco.") (Jacobs, 1995, 50,81-82.)

- **Imperatives** do not require a subject to be considered a clause as in: "Talk to me people!"

- In a sentence that has a subject with only an auxiliary verb, do not count the subject and verb as a separate clause. (e.g., "Cecilia is sad and her mother is too.") (Polio, 1997, 138-139.)

Error

- Consider that the text is a transcription of speech samples; therefore, do not count as errors any mechanical aspects of the text (e.g., capitalization, improper spelling, improper use of commas and periods)
- Consider following specific types of errors in counting (*Brown, Iwashita, & McNamara, 2005*):

a. Tense-marking errors:
   i. Omission of the past tense morpheme (i.e., -ed)
   ii. Overgeneralization of the past tense morpheme (i.e., -ed; buyed)
   iii. Use of the base form of an irregular verb/copular/auxiliary instead of a past tense verb/copular/auxiliary (e.g., “sink” for “sank,” “is” for “was,” “do” for “did”)
   iv. Use of the base form of a verb/copular/auxiliary where future tense is expected to be used (i.e., omission of auxiliary “will”)
   v. Use of the base form of a verb instead of the passive form (e.g., “pump” for “was pumped”)
   vi. Use of the base form of a verb instead of the progressive form (e.g., “increase” for “increasing”)
   vii. Use of the base form of a verb instead of a gerund or participle (e.g., “stop pump” for “stop pumping,” “reduce pumping by import water” for “reduce pumping by importing water”)
b. Third-person-singular verbs/copular:
   i. Omission of the third-person-singular morpheme (i.e., -s, -es)
   ii. Use of incorrect copular (e.g., “is” instead of “are”), irregular third-person-singular verbs (e.g., “have” instead of “has”)
c. Plural nouns:
   i. Omission of the plural (i.e., -s)
   ii. Use of a singular noun where a plural noun is required (e.g., “child” for children)
   iii. Overgeneralization of the plural; -s marker to an inappropriate context (e.g., At meal or breaks times students take the streets.)
d. Article use

i. Omission of indefinite and definite articles
ii. Incorrect use of articles (i.e., use of the definite article for the indefinite article and vice versa)

e. Prepositions
i. Use of an incorrect preposition
ii. Omission of a preposition
iii. Use of preposition in nonobligatory contexts