

DECISION MAKING WHILE SCORING EFL TAPE -MEDIATED SPEAKING TEST PERFORMANCE

Wang Haizhen
Soochow University

Abstract

This study investigates how raters make their scoring decisions when assessing tape -mediated speaking test performance . 24 Chinese EFLteachers were trained before scoring analytically five sample tapes selected from TEM4 -Oral , a national EFLspeaking test designed for college English major sophomores in China . The raters verbal reports concerning what they were thinking about while making their scoring decisions were audio -recorded and collected during and immediately after each assessment . Post scoring interviews were used as supplements to the probe of the scoring process . A qualitative analysis of the data showed that the rater tended to give weight to the content , to punish both grammar and pronunciation errors and to reward the use of impressive and uncommon words . Moreover , the whole decision -making process was proved to be cyclic in nature . A flow chart describing the cyclic process of hypothesis forming and testing was then proposed and discussed .

Key words

decision -making process ; raters ; tape -mediated speaking test ; TEM4

1 . Introduction

In many performance tests , including speaking and writing tests , test scores are reached subjectively using rating scale descriptors to guide the rater towards a score . Descriptions of a test taker s ability can then be produced by relating the score to the corresponding scale descriptors and the construct of language ability on which the rating scale is based . Yet , as commented by Connor Linton (1995 :763) , “if we do not know what raters are doing (and why they are doing it) , then we do not know what their ratings mean ” . Research on the rating (or scoring) process can direct diagnosis and improvement of rating scales (e. g. Hamp -Lyons 1991 ; Sakyi 2000 ; Vaughan 1991) , help interpret scores and report test results to score users in a valid , informed manner (e. g. Cumming et al . 2002 ; Orr 2002) , guide and monitor principled rater training (e. g. Weigle 1994) , and help understand the nature of the process (e. g. Milanovic et al . 1996) .

This empirical study , by collecting the raters verbal reports , investigates the scoring process in the Graded Test for English Majors — Band 4 Oral Test (hereafter TEM4 -Oral) , a large scale high stakes tape -mediated EFLspeaking test in China , with a view to better understanding how the raters made their scoring decisions .

2 . Literature Review

There have been a few exploratory studies on the process of assessing compositions since the 1970s ,

This paper is based on the author s PhD dissertation . Heartfelt thanks are given to Professor Wen Qiufang for her supervision .

when doubts were raised about the validity of existing holistic schemes for scoring compositions, both in English mother tongue and ESL/EFL contexts. In order to accumulate evidence to validate holistic scoring schemes, empirical studies were conducted regarding what trained raters attended to and how they applied the holistic scheme in the process of composition marking (e.g. Cumming 1990; Cumming et al. 2002; Huot 1993; Lumley 2002; Sakyi 2000; Vaughan 1991). Varied findings were yielded.

For example, Vaughan (1991) asked nine trained raters to read and rate six essays on a 6 point rating scale. The raters think aloud protocols were collected and analyzed. Results showed that, despite similar training, different raters focused on different essay elements and had characteristic styles of reading essays such as the "first impression dominates" approach, the single focus approach and the "two-category" strategy. Sakyi (2000) analyzed six experienced raters think aloud protocols collected when they independently read and scored 12 English essays written by first year university students. He found that the raters tended to depend upon certain dominant factors such as the presentation of ideas, length of the essay, and the presence or absence of grammatical errors. Variations in the raters attention were also reported in Cumming et al.'s (2002) study. Ten experienced ESL/EFL raters produced concurrent think-aloud protocols of their decision making while they rated randomly sequenced samples of 60 TOEFL essays. Results revealed that ESL/EFL raters attended more extensively to language than to rhetoric and ideas overall. In particular, the raters tended to devote greater attention to rhetoric and ideas in essays they rated high, and relatively more attention to language matters in essays they rated low.

In comparison to research on the process of written language assessment, the process of spoken language assessment has been underexplored. The few studies only focused on the rating process of oral interviews, and investigated in particular what raters attended to and how they used rating criteria (e.g. Nakamura 1996; Orr 2002). So far there have been no empirical studies of the process of assessing tape-mediated speaking performance. Little is known about what is happening in the raters mind, what aspect(s) of language they pay attention to and how they make their scoring decisions when assessing tape-mediated speaking performance.

3. Research Design

3.1 TEM4-Oral

TEM4-Oral is a nation-wide tape-mediated EFL speaking test designed for college English major sophomores in China. The test takers elicited responses to three different tasks (i.e. Task I: Retelling a story; Task II: Talking on a given topic; Task III: Role playing) are first recorded onto tapes in language labs, and then collected and sent to the test center. The scoring procedures involve mainly the training session and the scoring session. In the training session, all the EFL teachers who are sent from colleges and universities in different parts of the country are systematically trained before they become accredited TEM4-Oral raters. Then the trained raters apply the given four-level analytic rating scale (with a 4 representing "excellent" and a 1 representing "fail") to each student's performance and assign scores. Each tape recording is scored independently by two raters for five assessment categories, including contents of the three tasks respectively, pronunciation and intonation, and grammar and vocabulary. If the two raters fail to show adequate agreement, the sample of performance will be assessed by a third rater.

3.2 Research questions

This study aims to investigate how raters make their scoring decisions when assessing tape-mediated EFL speaking test performance. It attempts to address the following research questions:

- 1) How do the raters handle the relationship of different assessment categories on an analytic scale?
- 2) On what evidence do the raters base their scoring judgments?
- 3) What processes do the raters go through when making their scoring decisions?

3.3 Subjects

24 EFL teachers from eleven different universities in China participated in this study. Their EFL teaching experience ranged from one year to twenty-one years and their TEM4-Oral rating experience ranged from zero (11 raters) to nine years. Four were male and twenty were female. The gender

proportion and the diverse nature of the subjects' backgrounds were comparable to the distribution of the raters who participated in the actual TEM4 -Oral scoring session .

3.4 Instruments

Five sample tapes were selected and used for this study , which were representative of performance at different levels in the scale (with two tapes representing " pass "). A training package was used for the training session , which contained a sheet of Instructions for the Rater , a sheet of rating scale and marking specifications and a marking form sheet .

A profile questionnaire was designed to collect each rater's relevant background variables , such as gender , teaching experiences , related rating experiences and perceptions of EFL speaking ability .

Concurrent verbalization , " think aloud " or " talk aloud " (Ericsson & Simon 1993) , which would " happen at the same time that the information is being attended to " (Goh 1998 :127) , was used as optional in this study , due to the difficulty for the rater to verbalize thoughts concurrently while the attention was on listening and making constant judgments . Immediate retrospective verbalization was used to generate or elaborate thoughts immediately when each scoring task was completed and when there remained in short-term memory " retrieval cues that allowed effective retrieval of a subset of the sequence of thought " (Ericsson & Simon 1993 :149) .

The method of stimulated recall (Gass & Mackey 2000) was used as a complement when the subject was unable to self-report sufficiently the heeded information that was going on in her mind while scoring , or when the researcher as nonparticipant observed some points that needed clarification . In either case , the subject was asked to recall her thoughts with the support of a replay of the tape recording .

Semi-structured interviews were given to the subjects after they finished the scoring and reporting sessions . These were aimed at " tracing a process of how a particular decision was made " (Rubin & Rubin 1995 :29-30) . The major questions that the researcher intended to probe and follow up concerned what criteria were used for speech quality judgments and how a scoring decision was made .

3.5 Data collection

The data collection went through four stages : preparation , training , scoring and reporting , and post-scoring interview . Each session was carried out between one subject and the researcher herself .

In the preparation session , questionnaire responses were collected , and the subjects were given instructions about the data collection procedures .

In the training session , each subject was trained , following the basic procedures that are regularly used for the actual TEM4 -Oral training sessions , which include a study of the rating scale and marking specifications and a test marking . What made the current training different was that only the part of the recording concerning the monologic or narrative task (Task II) was heard and assessed .

After the training session , the subject was ready for scoring and reporting . Each rater listened to randomly sequenced samples of five speech recordings , and assigned separate scores for three categories of Content , Pronunciation and Intonation , and Grammar and Vocabulary . While she was making scoring decisions , the rater was asked to verbalize , concurrently and /or retrospectively , what came to the mind . When she was unable to verbalize thoughts at length while or immediately after scoring or when the researcher needed to clarify some specific points , the rater was asked to listen to the tape recording or fragments of the recording for the second time and to recall what she had been thinking about while scoring during the first listening .

When each rater finished scoring all the five sample tapes , she conversed immediately with the researcher for a post-scoring interview . Guided by the interview schedule and the observation notes , the researcher asked concerned questions and probed into the scoring process .

Both the scoring and reporting session and the interview session were audio-recorded . Each individual recording lasted from 71 to 146 minutes . They added up to about 2424 minutes in length .

3.6 Data analysis

The audio recordings were first transcribed and then developed into verbal report protocols. The verbal protocols were extensive, ranging in length from six to nine typed pages per rater, and amounted in the aggregate to 168 typed pages. Data analysis attempted less to quantify instances of each rating behavior than to identify what range of behavior could be observed and to portray a general scoring process that was followed by the raters. The data analysis, in other words, was qualitative rather than quantitative.

As the researcher was present at each data collection session, rough transcriptions and interpretations were made immediately after each session. Then these initial perceptions and interpretations contributed to the development of the system of generating themes, which were checked against the interpretation of new data collected. What the raters reported in the interviews was cross checked in their verbal protocols to see whether what they said consisted with what they did. Impressionistic interpretations of patterns and variations in the protocols were finally formed after revisions.

4. Results and Discussion

4.1 The relationship between three assessment categories

For each tape-mediated performance, there were three assessment categories for the raters to address on an analytic scale, that is, Content, Pronunciation and Intonation (P & I), and Grammar and Vocabulary (G & V). The researcher wanted to know whether the raters assessed each category independently, as required by the test developers, and if not, what aspect of language they tended to value and why.

Data analysis showed that only five raters reported assessing each category independently without giving extra weight to one particular category, and that a great majority of raters tended to attach special importance to either content or pronunciation and vocabulary. Table 1 shows the number of raters who handled the relationship between the three categories differently.

Table 1. The relationship between three assessment categories

	Balanced	Imbalanced	
		Content dominates	P & I dominates
N of raters	5	16	3
A brief description	The raters assessed each category independently, without giving extra weight to any particular category.	Of the three categories, the raters stressed content.	Of the three categories, the raters stressed pronunciation and intonation.

4.1.1 A balanced relationship

Five raters reported having assessed each category independently. Two of them were experienced raters (with rating experience 3 years) and the other three were inexperienced raters (with 0-1 years rating experience).

The two experienced raters behaved similarly in the scoring process. While listening to each sample of performance, both of them concentrated on recording and gathering evidence, based on which they made their scoring judgments. They reported that they took notes in order to ensure the objectivity of the test scores. As experienced raters, they were clearly aware that they were expected to treat different categories separately and they intended to do as required.

The three inexperienced raters seemed less confident and more cautious about their scoring judgments. They reported that their scoring was strictly criterion-referenced. While listening to each tape-recorded performance, they referred frequently to the level descriptors in the rating scale, and tried

to match the varied performance features with the descriptions . They tended not to think about the relationship between the categories .

4.1.2 An imbalanced relationship

Nineteen raters (80 %) reported having given special weight to one particular category .

Sixteen of them believed that in oral communication the successful transmission of the ideas was the most important of all and that the content was more important than the language form . Some of these raters stated frankly that the assessment of the other two categories was influenced by the assessment of content . For example , one rater gave priority to communicative ability over language ability and remarked that the test taker would not pass the exam if she failed to accomplish the communicative task .

Three raters were inclined to give more weight to the category of pronunciation and intonation while making scoring decisions . In the interviews they explained that the students were not very distinctive from each other in their presentation of the content and as a result the category of content lost its distinguishing power . Furthermore , one rater reported that her experience in teaching Spoken English and tutoring students for speaking contests drove her to be rigid with pronunciation and intonation . A second rater , in response to the pre scoring questionnaire item as regards what composed EFL learners speaking ability , ranked pronunciation and intonation first , followed by grammar and vocabulary . Apparently her view about EFLs speaking ability was consciously or unconsciously brought into her decision -making process .

4.1.3 Discussion

According to the scoring rubrics , the raters were supposed to assess each of the three categories separately . Then why did the majority of the raters give special weight to one particular category ? The answer to this question should be traced back to notions of linguistic competence and performance .

Linguistic competence normally refers to knowledge of the formal linguistic system as opposed to the application of this knowledge in language performance (Chomsky 1965). For language testing , a distinction is made between tests of " knowledge " and tests of " performance . " While in " knowledge " tests , proficiency in language is determined on the basis of knowing about the language without the need to prove it in communicative situations , " performance " tests require the test taker to apply the knowledge by actually using the language in such situations (Shohamy 1983 :528).

TEM4 oral is a syllabus based performance test . As a performance test , it tests not only language knowledge but also the " ability for use " (Hymes 1967). As a syllabus based test , it examines whether the test takers are " capable of communicating with native English speakers in social communications " (Syllabus for College English Majors , 2000 :8). In either case the ability to use the language communicatively becomes the center . The communicative nature of the test explains why sixteen raters in this study attached importance to meaning or content in the communicative task .

In addition , when the raters listened to the tape recorded performance and made scoring judgments , they played the dual roles of both the judge and the listener . As a listener , the rater would imagine having communication with the speaker , " as if he were telling a story to me face to face " (Feil , T3 immediate retrospection) . From the listener's standpoint , the rater was concerned about whether the talk was comprehensible and whether the communication proceeded successfully ; from the judge's standpoint , the rater judged whether the communicative task was accomplished in an effective way . It would be a natural tendency for the raters to be meaning driven when assessing a communicative language test .

Three raters were found to give weight to the formal aspects of pronunciation and intonation . Further analysis of the interview data revealed that this scoring tendency was affected by the lack of discriminating power of the test takers content , the raters previous teaching experiences , and for the raters beliefs about what was important in EFL learners speaking ability . Nevertheless , these raters did not deny the importance of content .

In conclusion , there was a prevailing tendency for the raters to give special weight to the category of content . The importance they attached to the content reflected the importance of comprehensibility as well as meaning transmission in communication . It is suggested that the test developers take this finding

into account and consider what instructions they give to the raters in the training session .

4.2 Evidence for scoring judgments

The raters' scoring judgments were dependent on the evidence heard and accumulated . The researcher wanted to know what types of evidence the raters gathered . The verbal protocol analysis revealed that the raters tended to search for different types of evidence for different assessment categories . When assessing Vocabulary , most raters listened for positive evidence ; when assessing both Pronunciation and Intonation and Grammar , most raters depended on negative evidence ; when assessing Content , all the raters turned to a combination of both negative and positive evidence . Table 2 presents the number of raters who relied on different types of evidence for assessing different categories .

Table 2 . Evidence for scoring judgments

Category	Evidence	Evidence		
		Negative	Positive	Negative + Positive
Content				24
P & I		21		3
G & V	Grammar	24		
	Vocabulary		20	4

4.2.1 Negative evidence driven

On the whole , the raters attended to negative evidence more often than positive one . All the twenty-four raters in this study were driven by negative evidence when assessing Grammar , and twenty-one raters relied on negative evidence for the assessment of Pronunciation and Intonation . As a consequence they adopted a subtractive scoring approach for both Grammar and Pronunciation .

As EFL teachers these raters appeared to be very sensitive to both grammar and pronunciation errors , especially when the errors appeared frequently within the speech and across the speeches . Maybe it is true that errors were more salient and caught the raters' attention , as commented by Mei (Excerpt 1) .

我习惯于减分。如果用对了 ,就像在手上滴的水 融合了 ,没有留下印迹 ;如果是错的 ,它不融 ,也就容易注意。(I m accustomed to making deductions . If the language form is used correctly , it will melt away like a drop of water in your palm , leaving no trace . If , on the other hand , the language form is used wrongly , it will not melt away and will easily catch your attention .)
(Excerpt 1 : Mei , interview)

4.2.2 Positive evidence driven

Twenty raters were found to be positive evidence driven when assessing Vocabulary . Only four raters made both positive and negative comments on Vocabulary in their verbal reports .

For vocabulary , the majority of the raters used an additive scoring approach . However , the positive comments that the raters made on vocabulary were few in number . It seemed that the raters did not pay much attention to vocabulary until the test taker occasionally came up with some impressive words , such as notice and hesitate in T3's speech , which "lightened my eyes momentarily" (Jun , immediate retrospection) , or reluctant , incidentally and wards in T1's speech , which were "beyond the level of average students" (Lan , immediate retrospection) . "Big" or "uncommon" or "impressive" words that occurred rarely and unexpectedly in the speech seemed to catch the raters' attention more easily . Excerpt 2 shows Fei's comment on the salience feature of good words , which were in her words "like fish jumping out of the ocean one after another" .

多数学生停留在一个等级上.....词汇也是这样 ,多数情况下就像在大海上风平浪静的。如果一个好的孩子突然之间词汇用得特别丰富 ,一个词一个词蹦出来 ,一条鱼一条鱼蹦出来..... 就会有加分考虑。(Most students remained at a similar level . . . The same happened with vocabulary . In most situations the vocabulary production remained peaceful and calm like an ocean . If suddenly a good

student used a wide range of vocabulary and produced one good word after another ,like fish jumping out of the ocean one after another . . .I would consider rewarding her .)

(Excerpt 2 :Fei ,interview)

4.2.3 Combining negative and positive evidence

All the twenty four raters considered both negative and positive evidence when assessing Content ; three raters attended to both negative and positive evidence while judging Pronunciation and Intonation ; four raters made both positive and negative comments on Vocabulary in their verbal reports .

Firstly ,all the raters combined both negative and positive evidence for the assessment of content . For different criteria they resorted to different types of evidence . When the raters were applying such criteria as content relevance ,coherence and story completeness ,they were inclined to depend on negative evidence . When the raters were considering such criteria as content sufficiency ,novelty and vividness in narration ,they were more likely to gather positive evidence .

Secondly ,three raters (Fei , Shan , Xian) considered both negative and positive aspects of pronunciation and intonation . Shan said in the interview that as she taught Phonetics she tended to be harsh about students pronunciation and intonation . She not only applied the criteria of accuracy ,fluency and naturalness in “score deduction ” but also looked for something “deserving rewards ”such as good use of rhythm ,linking and sense groups . Similar to Shan ,Xian not only gathered evidence of errors in pronunciation but also rewarded the students who used “ weak forms ”and “rising and falling tones ”in the talk . The third rater (Fei)regarded Pronunciation and Intonation as the least important category ,and she formed a holistic impression based on evidence from both sides .

Thirdly ,four raters made both negative and positive comments on Vocabulary . Two of them belonged to the type of raters who kept busy searching for and recording errors during listening and made deductions after listening . In rare cases they commented on the good use of words or phrases . The other two raters paid attention to the incorrect and inappropriate use of words and expressions ,and meanwhile they kept a sharp ear for the rare appearance of good lexis .

4.2.4 Discussion

As noted above ,there was a tendency for the raters to turn to a different type of evidence for the assessment of a different category . We may wonder why .

First ,as EFL teachers the raters appeared to be very critical about the formal aspects of the language . They were not very tolerant of errors in the speeches . It is likely that the raters severity about errors was determined by their experiences in EFL teaching and by their being non native English speakers . As shown in previous studies ,teachers were more critical and tended to be stricter on linguistic items such as grammar ,vocabulary and expression than non teacher raters ,and NNS (non native speaker)raters tended to have a more negative view than NS (native speaker)raters ,seeing any deviance from the standard as errors ,and marking students down accordingly (e. g. Galloway 1980 ;Brown 1995) .

Second ,it is interesting that the raters tended to look for errors when assessing grammar and pronunciation but looked for something to be rewarded when assessing vocabulary . If we ask why they looked for evidence ,we will realize that they were indeed looking for discriminating features in the speech . When assessing a tape mediated performance ,with the recording playing on and on ,it is almost impossible for the raters to listen to the speech word by word or even sentence by sentence . The raters have to concentrate their attention on the most helpful and salient features in the performance ,features that help them to distinguish students at different levels . In the performance presented to the raters in this study ,it was found that the EFL learners at the tertiary level either made few errors in the language form or chose not to take risks in a test situation by using the words they were not certain about . Consequently few uncommon or sophisticated words were used in the performance . In this case ,errors in either grammar or pronunciation became salient and the rare occurrence of good impressive words became noticeable . As a result ,the raters tended to be negative evidence driven for grammar and pronunciation , and positive evidence driven for vocabulary . We may imagine that if a great majority of the students made a large quantity of errors in their speeches ,the raters would pay more attention to the correct

forms. Moreover, it is an underlying tendency for people to take short cuts and for raters to find a more efficient and effective way of fulfilling the scoring task. Looking for discriminating features helps solve the problem of how to discriminate between students in the shortest time possible.

Third, the tendency to look for discriminating features also applied to the assessment of content. In the performance, most students were found to be able to tell a relevant, complete story logically, and few students displayed the features of sufficiency or novelty or vividness. Correspondingly, when considering the relevance, completeness and logic of the content, the raters tended to look for negative evidence; when considering content sufficiency or topic novelty or vividness, the raters were more likely to look for positive evidence.

4.3 The cyclicity of the scoring process

This section addresses the third research question: "What processes do the raters go through when making their scoring decisions?" In particular, the researcher wanted to know how the raters assessed three different categories simultaneously and what process was going on when assessing the tape-mediated performance. Data analysis revealed that on the whole the raters adopted a cyclic way of assessing three different categories and of assessing each individual category.

4.3.1 A cyclic process in assessing three categories

The raters were found to assess the three categories in a cyclic process. They tended to focus their attention on one category at a time, form a score hypothesis for this category, and shift their focus of attention to the next category once the hypothesis about the first category became relatively steady. When they were assessing the second category, they were simultaneously attending to features relating to the other two categories, and were ready to return to the first category and modify the hypothesis when additional evidence was detected. In this way, the raters assessed the three categories one by one in a sequence, and on the other hand they paid attention to varied features relating to the three categories simultaneously and went through the assessment process of every category in a recursive or cyclic process.

Table 3 illustrates an example of this cyclic process of assessing three different categories. Excerpt 3 was selected from Yan's immediate retrospection while assessing T5. Yan chose to pause the tape recorder whenever she felt that something came to her mind and then verbalized her thoughts immediately within seconds. Then she pushed the play button and listened on. According to the times when Yan paused the recorder and verbalized her thoughts, the whole excerpt was broken into eight episodes in Table 3. The sequence of these episodes followed the original sequence of Yan's verbalizations. In Table 3, the episode number is listed in the left column, Yan's verbalizations are presented in the middle column, and the author's explanatory notes are offered in the right column.

Table 3. Illustrating a cyclic process in assessing three different categories

Epd	Excerpt 3	Notes
1	刚开始第一句话她语音后面加了个元音 / / , /h um / [home]。这是中国学生容易犯的错误。语音语调首先最高不会达到80分。 (In the very beginning utterance she adds a vowel / / after pronouncing [home], /h um /. This is an error frequently committed by EFL Chinese students. Pronunciation & Intonation will not reach a maximum of 80 points.)	Primary A :P & ; Initial H1 : <80
2	value things 应该是 valuable / ... fell into horrible , 这是个 chunk 错误。 (value things , it should be valuable . / ... fell into horrible . This is a chunk error .)	Primary A :G & V ; Initial H2 :negative
3	到这个地方她要讲的故事我也能猜出来了 ,就是有人路遭抢劫 ,有人拔刀相助。 (By this point I can guess what story she is going to tell . It s about someone who was robbed on the way , and someone else helped despite danger to himself .)	Primary A :Content ; Initial H3 :average (unoriginal)

Epd	Excerpt 3	Notes
4	她的语法在tense 上把握还是蛮好的 ,听了几个pasttense 都是用的正确的。(Her grammar in tense is good . She used past tense correctly several times .)	Primary A :G &V Modify H2 (H2) :positive
5	又一个 / / /b t / [but] 进一步印证了我的想法 ,语音语调只能给70 分。(Another / / , /b t / [but] . This further confirms my judgment . Pronunciation & Intonation can only reach 70 points .)	Primary A :P &I ; Modify H1 (H1) :70
6	到现在为止没听到特别好的词汇 ,就是很一般。 [70 分] (Up to now I haven t heard any particularly good words . Vocabulary is average .)	Primary A :G &V ; Modify H2 (H2) :average Finalize H2 :Score =70
7	hospital 重音有问题 ,语音语调65 分吧。(An error in the stress on hospital . Pronunciation & Intonation is now 65 points .)	Primary A :P &I ; Modify H1 (H1) :65 Finalize H1 :Score =65
8	内容她讲的是个完整的故事 ,但这个故事比较老套 ,大家都能猜到.....我给70 分。(The story is complete ,but unoriginal in topic . Everyone can guess the ending I assign 70 points .)	Primary A :Content ; Modify H3 (H3) :average Finalize H3 :Score =70

Epd =Episode ;A =Attention ;H1 =hypothesis of P &I ;H2 = hypothesis of G &V ;H3 = hypothesis of Content

As Table 3 shows ,Yan s primary attention was focused on one category at one time and went in a cycle for features related to the three separate categories . In the first three episodes (i. e. the initial four sentences in the speech) , Yan assessed the speaker in a sequence of Pronunciation and Intonation first , Grammar and Vocabulary second and Content third ,and accordingly built three separate scoring hypotheses (H1 ,H2 ,H3) . Then Yan s focus of attention shifted to Grammar and Vocabulary first (Episode 4)and then to Pronunciation and Intonation (Episode 5) before it shifted back to Grammar and Vocabulary (Episode 6). Thus Yan s primary attention went in a cycle for salient features related to the three categories . Finally ,after one or two modifications (e. g. H1 ,H1) ,Yan finalized her scoring decisions (Episode 6 ,7 ,8) . Yan s scoring process was cyclic in nature . Table 4 gives a better illustration .

Table 4 . Illustrating a cyclic process in assessing individual categories

P &I	G &V	Content
Episode 1 Initial H (H1) : <80	Episode 2 Initial H (H2) :negative	Episode 3 Initial H (H3) :average
Episode 5 Modify H (H1) :70	Episode 4 Modify H (H2) :positive	Episode 6 Modify H (H2) :average Finalize H2 :Score =70
Episode 7 Modify H (H1) :65 Finalize H1 :Score =65	Episode 8 Modify H (H3) :average Finalize H3 :Score =70	

In Table 4 ,if we draw a line to connect the episodes from 1 to 8 successively ,we will find that this line is not a straight line but a zigzagging line ().It indicates that the scoring process for the three categories is not linear but rather cyclic in nature .

4.3.2 A cyclic process in assessing individual categories

For each assessment category ,the raters also made judgments in a cyclic process ,continuously forming preliminary scoring judgments ,attending to specific features ,confirming or revising the preliminary judgment ,and listening for more specifics and making more modifications ,and so on . To generalize ,the scoring process for each category was found to be a dynamic process of sampling and scoring ,and resampling and rescoreing ,before the decision was finalized and a final score was assigned . The whole score decision-making process can be described as a hypothesis forming and testing process (Figure 1).

In the flow chart ,an oval means a starting point or an ending point (terminal) ;a rectangle means an operation ;and a diamond means a decision to be made between N and Y . Nstands for “ No ” ,whereas Y stands for “ Yes ” . The question mark (?)signifies a question to be answered or a decision to be made , while arrows indicate the flow of each operation /decision .In Figure 1 ,the flow chart starts with start , which indicates that the play button of the recorder is pushed and the tape recording starts playing ; sampling and resampling refer to the rater listening to a recording ,either intentionally sampling the recorded performance for scoring judgments or making natural response to specific features that emerge in the performance ;scoring and rescoreing refer to the rater making decisions and /or modifications on the score ,based on information gathered from sampling and resampling ;finalized refers to the raters scoring decision being finalized or the tape recording playing to the end ;and final score refers to a final score being assigned ,which signifies the terminal of the flow chart .

Figure 1 . A flow chart of hypothesis forming and testing process

As depicted by this flow chart ,when the rater was listening to the speech ,he or she started gathering information and formed a rough impression of the test takers relative position on the rating scale . Based on the initialtwo or three utterances ,the rater formed an initialscore hypothesis . Then the rater went on listening ,sampling and making judgments ,in an attempt to justify the initial score

hypothesis with more concrete evidence .

If the rater noticed more speech features that matched in every way the descriptions of the hypothesized level or score ,the rater would rest assured that the initial hypothesis had been confirmed . What the rater did next was to finalize the score and end the scoring process . The Y arrow directions in the flow chart show this linear sequence of forming an initial scoring hypothesis ,confirming and finalizing the hypothesis .

However ,in most cases the initial score hypothesis needed to be modified several times ,when more errors were detected ,when one or two impressive words were heard ,or when the pros and cons of varied speech features were traded off . In this case ,the increasing evidence did not show a close match of the performance with the descriptions of the particular level or score hypothesized ,therefore calling for a modification of the initial hypothesis . So the rater repeated the cycle of (re)sampling the performance and (re)scoring or modifying the hypothesis until in the end a more satisfactory score was finalized . The N arrow directions in the flow chart show this cyclic process of hypothesis forming (scoring) ,modifying and reforming (rescoring) . . . and finalizing .

To illustrate this hypothesis forming and testing process ,Yan's verbalizations while assessing T5 (Excerpt 3) , which were broken down into eight episodes and presented first in Table 3 and then reorganized in Table 4 ,are reanalyzed here .

In Table 4 ,the scoring process of each category is displayed in columns ,and the related episodes are put into the corresponding column vertically . If we follow the episodes related to each category vertically ,we are tracing a process of assessing each individual category ,that is ,a hypothesis forming , modifying and finalizing process . Take Yan's assessing of Pronunciation and Intonation for a closer examination . The related episodes in the left column are Episodes 1 ,5 and 7 . From Episode 1 ,we see that Yan's initial score hypothesis (H1 : <80) was formed at the very beginning of the speech ,when she detected a mispronunciation . Then her focus of attention was shifted to other categories (Episodes 2 ,3 , 4) before it returned to Pronunciation and Intonation for the first time (Episode 5) . This time based on a similar mispronunciation detected ,Yan confirmed her judgment and modified the initial hypothesis (H1 : 70) . Soon when Yan noticed an error in stress ,she shifted her primary attention back to Pronunciation and Intonation for the second time (Episode 7) ,and modified the scoring hypothesis again (H1 : 65) . It is assumed that Yan would continue modifying the scoring hypothesis if the speech lasted longer and /or if more negative or positive evidence was gathered . For this particular test taker (T5) the score of 65 points was the final score that Yan assigned .

4.3.3 Discussion

As discussed above ,there is a big cyclic process in which the raters assess the three different categories ,and in which smaller cyclic processes go on for the assessment of each individual category . By focusing their primary attention cyclically on the three categories and by forming , modifying and finalizing scoring hypotheses cyclically , the raters were actively constructing scores . The whole assessment became a score construction activity .

The assessment of speaking performance ,like writing assessment ,is an example of "an ill structured task " (DeRemer 1998 :14) . Despite standardized training procedures ,there is no standard solution procedure for the speech assessment . The raters in this study were given scoring rubrics including the rating scale and marking specifications for the narrative task ,and benchmark tapes for test marking . They were trained to reach a common understanding of the descriptors in the rating scale ,to have an understanding of what type of performance matched a typical level on the scale and what specific performance features or aspects of aural text to assess . In contrast with the traditional view that the scoring process involved simply matching the speech sample with the descriptors on the scale and assigning a score accordingly (e.g. Shohamy 1988) ,the raters' scoring processes were found to be more complex , in which the raters were actively forming hypotheses ,modifying and finalizing hypotheses in a cycle . In their scoring processes ,the raters were neither given standard solution procedures to problems such as "What types of grammatical errors are grave errors ?" nor provided with a ready-made representation of the scoring activity . The raters did not act passively like machines or instruments . On the contrary ,they

were obliged to respond actively and develop their own plans of action. In this sense, language assessment becomes a constructive activity, in which the rater, not the rating scale, lies at the center. It is the rater who decides what initial scoring hypothesis to build, when to modify the hypothesis and when to finalize the scoring decision, etc. It is the rater who decides which features or criteria in the scale to pay attention to, how to arbitrate between uncertainties and scale descriptors, and how to justify the final score decisions in terms of scoring rubrics and expertise judgments.

5. Conclusion

Major findings in this study can be summarized as follows. First, there was a tendency for the raters to give weight to the category of content when assessing a communicative task. Second, when assessing different categories, the raters tended to look for discriminating features that help distinguish students at different levels and thus depend on different types of evidence. In particular, the raters were negative-evidence driven when assessing the formal aspects of language (i.e. Grammar, Pronunciation and Intonation); were positive evidence driven when assessing Vocabulary; and they combined both negative and positive evidence when assessing Content. Third, the scoring process was a cyclic process in which the raters' attention was directed in a cycle to features relating to different categories. A hypothesis forming and testing flow chart was proposed to depict the cyclic decision-making process. It was also suggested that speech assessment involves a complex process in which the raters are actively constructing scores.

Notes

1. To preserve the confidentiality of individual raters, pseudonyms are used. T3 stands for Tape 3; immediate retrospection shows the origin of the quotation.
2. The original words were in Chinese; the English translation was the author's.

References

- Brown, A. 1995. The effect of rater variables in the development of an occupation specific language performance test. *Language Testing* 12 /1 :1-15.
- Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT press.
- Connor-Linton, J. 1995. Looking behind the curtain: What do L2 composition ratings really mean? *TESOL Quarterly* 29 :762-765.
- Cumming, A. 1990. Expertise in evaluating second language compositions. *Language Testing* 7 :31-51.
- Cumming, A., R. Kantor & D. E. Powers. 2002. Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal* 86 :67-96.
- DeRemer, M. L. 1998. Writing assessment: Raters' elaboration of the rating task. *Assessing Writing* 5 /1 :7-29.
- Ericsson, K. & H. Simon. 1993. *Protocol Analysis: Verbal Reports as Data* (2nd ed.). Cambridge, MA: MIT Press.
- Galloway, V. B. 1980. Perceptions of the communication efforts of American students of Spanish. *The Modern Language Journal* 64 :428-433.
- Gass, S. & A. Mackey. 2000. *Stimulated Recall Methodology in Second Language Research*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.
- Goh, C. M. 1998. How ESL learners with different listening abilities use comprehension strategies and tactics. *Language Teaching Research* 2 :124-147.
- Hamp-Lyons, L. 1991. Reconstructing "Academic writing proficiency." In L. Hamp-Lyons (ed.). *Assessing Second Language Writing in Academic Contexts*, pp. 127-153. Norwood, NJ: Ablex.
- Huot, B. 1993. The influence of holistic scoring procedures on reading and rating student essays. In M. Williamson & B. Huot (eds.). *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations*, pp. 206-236. Cresskill, NJ: Hampton Press.
- Hymes, D. H. 1967. Models of the interaction of language and social setting. *Journal of Social Issues* 23 :8-38.
- Lumley, T. 2002. Assessment criteria in a large scale writing test: What do they really mean to the raters? *Language Testing* 19 :246-276.

(Continued on p. 16 ...)