Teachers' and Non-Teachers' Perceptions of a Chinese Learner's Oral Performances

Guangyan Chen
Texas Christian University

Abstract

This study explores whether teachers differ from non-teachers (naïve native speakers) in their perceptions of a learner's oral performances. The oral language performances of an American university student in a fourth-year Chinese class were videotaped. Descriptive items were generated for use in evaluating the learner's performances. Then 343 Chinese teachers and non-teachers used these items to evaluate the performances. The data were analyzed by means of **exploratory** factor analysis and analysis of variance. No significant differences were found between teachers and non-teachers in their rating criterion patterns. This result suggests that experience as a teacher does not necessarily out weight culturally-influenced perceptions. The implication is that experts in language assessment should make studying the underlying perceptual judgments of native speakers in the assessment of oral performance a top priority.

Introduction

The default goal of learning a foreign language is to communicate effectively with native speakers. Learners are the direct consumers of a foreign language program. Native speakers of a target language are the recipients of a foreign language program by communicating with these learners. Therefore, they naturally become the group of people whose perceptions and standards can work as the norm to evaluate these learners' communicative capacity. However, their assessment is regarded as naïve, since they are not trained teachers, nor are they trained in assessment or have descriptive criteria against which to judge the performances. Many studies, such as the studies of Elder (1993) and Shohamy et al. (1992) suggest that non-trained raters' judgments should be handled carefully owing to rating inconsistency.

In language assessment practice, teachers and trained raters are usually the population to conduct assessment practice. The problem is teachers' evaluation is not without questions. As Shohamy et al. (1992: 28) pointed out, "The professional background of the rater may affect rater reliability. Teachers, for example, may be influenced by instructional goals and may emphasize particular components of the written sample." Therefore, the relevant question is whether raters of different backgrounds, such as teachers vs. linguistically naïve native speakers, differ in their ratings.

The purpose of this study is to test whether teaching background can influence native speakers' perception in rating oral performances. The hypothetical assumption of this study is that native speakers evaluate these learners' performances according to their somewhat ingrained perceptions. These perceptions are hard to change. As Brown states, "Raters appear to have inbuilt perceptions of what is acceptable to them and these perceptions are formed to some extent by their previous experience. It appears that even the explicitness of the descriptors and the standardization that takes place in a training session cannot remove these differences." (Brown: 1995, 13) Therefore, the study intends to test the hypothesis as to whether native speakers' inbuilt cultural factors will dilute the effect of teaching experiences on their perceptions in rating performances. Namely,

whether teachers and non-teachers' perceptions differ in rating oral performances.

If the hypothesis is true, namely, naïve native speakers and teachers perceive language learner performances in a similar way, then the implication is that experts in language assessment should make studying the underlying perceptual judgments of native speakers in the assessment of oral performance a top priority. However there appears to be little empirical evidence to suggest that these assumptions about inbuilt perceptions hold true. This study fills the gap.

Many studies have explored the differences and similarities between native speakers' and nonnative speakers' evaluation practices (Galloway, 1980; Fayer and Krasinski, 1987; Brown, 1995; Shi, 2001; Kim, 2009). However, there has been little research about whether teachers and non-teacher raters (a term that refers here to linguistically naïve native speakers) give similar global and/or analytic ratings to the same speech sample. Only one study has explored whether teachers' and non-teachers' rating criterion patterns are similar or not (Hadden, 1983; 1991). Hadden concluded that non-teachers rate students' linguistic ability significantly higher than ESL teachers do. However, the overall rating patterns of non-teachers and teachers are similar.

A few other studies focus on comparing ratings of teachers and non-teachers in terms of specific rating dimensions, such as grammar, vocabulary, appropriateness, fluency, and content. Some of these studies found that teachers tended to be harsher than non-teachers (Galloway, 1977, 1980; Okamura, 1995; Kim 2009). Kim found that teachers (both native and non-native) are overall more severe in assessment than non-teachers. Okamura (1995) also found that teachers tend to be more critical than non-teachers in most rating dimensions. Galloway (1977; 1980) reported that teachers are more critical, particularly of pronunciation and rate of speech. Some studies, such as Chalhoub-Deville (1995), compared the ratings from three different rater groups found that teachers paid more attention to the creativity and adequacy of information than to linguistic features. However, other studies, such as Shohamy et al. (1992), found no differences between teachers and non-teachers, although trained and

non-trained raters differed. This result is consistent with Barnwell's (1989) findings that native speakers were consistently stricter in their evaluation than was an ACTFL trained rater.

The research described above provides scanty and inconsistent information about comparisons between teacher and non-teacher raters regarding their rating performances. Within this context, the present study explores and contrasts teachers and non-teachers (naïve native speakers of Chinese) in their perceptions of a Chinese learner's performance. The purpose is to explore whether teachers and non-teachers rate an oral performance using similar or different rating criterion pattern. For the rating criteria shared by the two groups, the question is whether one group is more critical than the other in rating performance.

By exploring and contrasting teachers' and non-teachers' rating criterion patterns, the study evaluates the construct validity of the ratings of teachers and non-teachers in the assessment of oral performance. Thus, the current study provides empirical evidence as to whether experts in language assessment should make studying the underlying perceptual judgments of native speakers in the assessment of oral performance a top priority.

Methodology

The general procedures for conducting the study are as follows. First, an American learner of Chinese was videotaped while speaking and the video was put online. Second, evaluation items relevant to the video were generated to evaluate the learner's performances. Then native speakers of Chinese (teachers and non-teachers) went online and evaluated the learner's performances using these items. Finally, two statistical procedures, Exploratory Factor Analysis (EFA) and Analysis of Variance (ANOVA) were run to analyze the data.

2.1 The Video-Based Oral Performance Sample

Oral performance video sample was created firstly as the target of assessment. The speaker in the video is an American learner of Chinese. She is a student in a college-level Chinese program. The student was learning Chinese in a Fourth-Year Chinese class during the experimental time. Ten students were originally selected from three levels: four students had finished one year of Chinese learning, three had finished two years, and three had finished three years. Then, one student was chosen from each level to represent performances and abilities at that level. The performances of these three students all received evaluation from native speakers of Chinese. However, only the evaluation of one student's performance is reported in this paper due to her proficiency level being in the middle among the three.

The video is five minutes long and covers seven topics. Several aspects were considered in the course of developing the video, such as the length of the video, the number and variety of topics queried, and the difficulty levels of the questions. The topics include: self-introduction, hometown description, economic crisis, family income, study abroad experiences, comparisons between Chinese culture and society in Chinese communities (Mainland China and Taiwan), and American culture and society. The difficulty of the topics ranged from personal questions to questions about social and cultural issues.

2.2 The Subjects

The respondents to the video are the subjects of this study. Three studies were conducted: a small pilot study, Study 1, and Study 2. Each study received responses from different subjects. The purpose of conducting three studies, rather than one, was to gradually refine observed items. In particular, EFA was used for item analysis. In performing EFA, it is crucially important to select observed items appropriately and carefully, as they predict latent factors (potential rating criteria) and have tremendous influences on the rating criterion pattern.

The pilot study was conducted in order to refine items and procedures prior to the two main studies. Subject samples in the pilot study were predominantly taken from Chinese students and scholars at two Midwestern schools. These data were not reported in this study. One month later Study 1 was conducted nationwide in order to further modify items. The subject samples were mostly collected from friends and colleagues. Overall, 133 subjects responded to the video. EFA was then run on these responses to determine the clusters of items. Items found not to cluster meaningfully with other items were revised or deleted. Two months later, Study 2 was conducted worldwide.
The subject sample for Study 2 was obtained from native speakers of Chinese with access to the following popular Chinese websites, such as http://www.mitbbs.com and http://www.huaren.us. The subjects were categorized[1] in terms of their background: teachers of Chinese in the U.S. sponsored by the office of the Chinese Language Council, teachers of Chinese who are on the mailing list of the Chinese Language Teachers Association in the year of 2010, those Chinese who work in international business companies in China and have had experience dealing with American learners of Chinese, and college students in the U.S. A total of 210 subjects responded to the video.

2.3 The Items Used to Evaluate Oral Performance

The initial generation of items used to examine respondents' perceptions of the oral performance sample was based on foreign language theory, published research (Palmer, 1973; Hadden, 1983, 1991), interviews with experts in Chinese language pedagogy, or some combination of these. The foreign language theory in this study refers to performed culture-based pedagogical philosophy, in which all dimensions of culture were considered when developing rating criteria (Walker, 2000; Walker and Noda, 2000).

---

[1] Before starting the assessment survey, the subjects need to fill in the form as to whether they teach Chinese before, where (K-12 or college; mainland China, the U.S., or other places) they teach, and how long they teach.

Table 1: The 17 Observed Items

The Items (Both in Chinese and in English)	The Expressions used in This Paper
1. 他/她□得很流利。He/she speaks fluently.	1 Fluent Speeches
2. 我□得他/她的交□水平比□高。**I think his/her communicative proficiency level is high.**	2 High Proficiency
3. 他/她表达得很清楚。**He/she expresses him/herself clearly**.	3 Clear Expressions
4. 他/她的想法和中国人差□很大。**His/her thought process is very different from that** *of* a native speaker of Chinese**.**	4 Significant Differences in Thought
5. 我□得他/她交流□很懂□□技巧。**I think he/she has an effective communication strategy.**	5 Communication Strategy
6. □些□□□他/她来□很容易回答。**These questions seem easy for him/her to answer.**	6 Questions Easy For Him/her to Answer
7. 他/她有很多□法□□。**He/she makes many grammatical errors.**	7 Many Grammatical Errors
8. 我□得他/她不太了解中国人的思想和行□。**I do not think he/she understands Chinese ways of thinking.**	8 Does Not Know Chinese Ways of Thinking
9. 我□得他/她不太理解要□的□□。**I think he/she does not understand the questions being asked.**	9 Does Not Understand the Interviewer
10. 我□得他/她是一个很□人喜□的人。 **I think he/she is a person that others would like.**	10 A Person Others Like to Deal with
11. 我□得他/她的回答，内容很丰富。 **I think the content of his/her answer is informative.**	11 Rich Content
12. 他/她□□造句很恰当。**He/she selects appropriate words and structures.**	12 Appropriateness in Choosing Words and Structures
13. 他/她很多□用□了。**He/she uses many words incorrectly.**	13 Wrong Word Usages
14. 他/她□中国的□知很多是□□的。 **He/she has misperceptions of Chinese cognition.**	14 Wrong Perceptions of Chinese Thinking
15. 我□得他/她听力不太好。 **I think his/her listening comprehension is poor.**	15 Poor Listening
16. 他/她□**的**□信息量很大。 **He/she delivers a good deal of information.**	16 Informative Speeches
17. 他/她的□音□□很得体。**His/her tone and intonation are appropriate.**	17 Appropriate Tone and Intonation

EFA provides another means of generating items beyond the initial set. According to the rule of generating observed items, three to

five conceptually related items are necessary to predict one latent factor. In running the data, the three to five items might load onto this latent factor. For example, "appropriateness" might be a latent factor in assessing learner performance. Therefore, the many items considered for use in this study included four items intended to predict the latent factor of appropriateness:

Item 1"He responds to questions appropriately (Ta yingdui hen deti.)."

Item 2 "He/she selects appropriate words and grammatical structures (Ta xuanci zaoju hen qiadang.)."

Item 3 "His/her tone and intonation are appropriate (Ta yuyin yudiao hen deti.)."

Item 4 "His/her body language is not appropriate (Ta zhiti yuyan butai deti.)."

After running EFA, there are two possible outcomes: The items either group together or are not highly intercorrelated. If one item does not cluster meaningfully with other items, then it is revised or deleted. In short, after running EFA, meaning-consistent items are kept, while other items are changed or deleted.

Finally, 17 items were kept and shared across Study 1 and Study 2. Responses to these 17 items were further analyzed, and were referred to here as the "S1&2 data." S1&2 data consist of the sums of individual responses within each study. In Study 1, the video received 123 responses, while in Study 2, the video received 210 responses. Thus the S1&2 data consisted of 343 responses. Responses could be combined across studies because the samples were mainly composed of different subjects. The combined data has much stronger reliability than the data collected separately in Studies 1 and 2.

2.4 The Five-Point Likert Rating Scale

All of the data were automatically collected from http://www.surveymonkey.com. A five-point Likert scale was used to measure extent of agreement with each item, with "strongly agree" receiving a score of 1 and "strongly disagree" receiving a score of 5. Therefore, for the items with positive descriptions, a low score represents a good evaluation. For the items with negative descriptions, a high score indicates a good evaluation.

2.5 Statistical Analysis

EFA

EFA is used to explore the possible underlying factor structure of a set of observed items without imposing a preconceived structure on the outcome (Child, 1990). By performing EFA, the underlying factor structure is identified. This procedure is often used to measure an ability or trait.

The first step in conducting EFA is to determine the number of factors. There are many methods for doing so. This study adopts the most frequently used and highly reliable techniques, the Cattell scree and eigenvalue-one methods. The Cattell scree is a graph of eigenvalues, derived from a preliminary factor analysis without rotation, which looks for meaningful clusters of items with fairly high factor loadings. After the initial step, the list of items for a factor is further examined and those items that do not have consistent meaning with the other items are eliminated. Finally, a label based on the common features of all items loaded on one factor is assigned to each factor. When labeling factors, I paid particular attention to those items with the highest factor loadings.

One type of oblique rotation, promax rotation, was used in this study. An oblique (rather than orthogonal) rotation was selected because oblique rotation assumes that factors correlate to each other. This view provides a more realistic solution in the field of rating criteria, given that rating criteria are often correlated in the real world.

ANOVA

The ANOVA technique was used to further compare teacher and non-teacher perceptions of oral assessment in terms of the identified factors in EFA. In the ANOVAs, there is one independent variable—the respondent group. This independent variable was fixed and active and has two levels: teacher versus non-teacher. Individual responses are nested within the two levels. The dependent variables were the rating criteria obtained from EFA of teachers' and non-teachers' responses to oral performance samples.

3. Results

This part reports the results of the separate EFAs for teachers' and non-teachers' responses, followed by a comparison of the responses between the two groups. Each EFA involved two steps. The first step was to determine the number of factors (i.e., rating criteria). The second step was to name these factors by analyzing a rotated pattern matrix of teachers' and non-teachers' responses.

3.1 EFA of Teachers' Responses

The scree plot of the teacher data shows a break between Factors 4 and 5, which suggests that four factors could be retained for this data. The eigenvalue-one rule also suggested that four factors could be retained, because Factors 1 through 4 had eigenvalues greater than 1, with the eigenvalues of Factors 5 and 6 close to the cut-off point. To avoid under factoring or over factoring, EFA was conducted with three, four, five, six, and seven factors, respectively. However, only the four-, five-, and six-factor solutions had consistent conceptual meanings and are reported here.

Table 2 shows the four- through six-factor solutions and internal consistency reliabilities. In this data, the factors, which had two items loaded on them, were retained because these items consistently combined together and constituted meaningful clusters. Analyses of the pattern matrices of all solutions revealed that the most detailed and meaningful clusters of items occurred when the number of factors retained was six. In fact, the difference between the five-factor solution and the six-factor solution was that F6[2] (Items 16, 11) and F1 (Items 5, 10, 6, 12, 17) in the six-factor solution were merged into F1 (Items 10, 5, 12, 16, 17, 11, 6) in the five-factor solution. The difference between the four-factor solution and the five-factor solution was that F5 (Items 7, 13) and F2 (Items 14, 4, 8) in the five-factor solution were merged into F2 (Items 4, 14, 8, 13, 7) in the four-factor solution. .

[2] F# stands for Factor #; for example, "F1" refers to Factor 1. Such representations apply to the whole paper.

Structure stability and consistency can be seen by comparing different factor solutions, as shown in Table 2.

Table 2: Factor Solutions of Teacher Responses

Factors	# of Items	Items*	Alpha
The four-factor solution			
#1º	7	10∞, 5, 12, 16, 17, 11, 6	.828
#2	5	4, 14, 8, 13, 7	.833
#3	3	1, 2, 3	.809
#4	2	9, 15	.584
The five-factor solution			
#1	7	10, 5, 12, 16, 17, 11, 6	.828
#2	3	14, 4, 8	.809
#3	3	1, 3, 2	.809
#4	2	9, 15	.584
#5	2	7, 13	.672
The six-factor solution			
#1	5	5, 10, 6, 12, 17	.783
#2	3	8, 14, 4	.809
#3	3	1, 3, 2	.809
#4	2	9, 15	.584
#5	2	7, 13	.672
#6	2	16, 11	.662

º #1 under the four-factor solution means Factor 1 in the four-factor solution.

* Items within each factor are listed in the order of the values of factor loadings, from the highest factor loading value to the lowest one. For example, # 4 under the four-factor solution has Items 9 and 15. The loading value of Item 9 is higher than that of Item 5. Therefore Item 9 is before Item 5.

∞ 10 represents Item 10 in the Study 1&2 data. The description of each item is listed in Table 1.

All of the representations apply to all similar tables in this paper.

Table 3: Rotated Pattern Matrix of Teacher Responses

Factors and Items	F1	F2	F3	F4	F5	F6
F1 Communication appropriateness						
5 Communication Strategy	**.874***	-.124		.102		-.139
10 A Person Others Like to Deal with	**.782**	-.284	-.140	.130	.130	.103
12 Appropriateness in Choosing Words and Structures	**.618**	.140	.159	-.233		
6 Questions Easy For Him/her to Answer	**.457**	.285			-.170	
17 Appropriate Tone and Intonation	**.424**					.201
11 Rich Content	***.387ø***	.134	.132			***.226***
F2 Chinese cognitive patterns						
8 Does Not Know Chinese Ways of Thinking		**.780**				-.252
14 Wrong Perceptions of Chinese Thinking		**.779**	-.101	.141	-.150	
4 Significant Differences in Thought	-.173	**.622**			.196	.183
13 Wrong Word Usages		***.425***		.242	***.322***	
F3 Communicative clarity						
1 Fluent Speech			**.856**	.139		
3 Clear Expressions		-.136	**.753**			
2 High Proficiency	.103∞		**.688**			
F4 Listening comprehension						
9 Does Not Understand the Interviewer	.106	.211		**.673**		
15 Poor Listening	-.117			**.594**		
F5 Language accuracy						
7 Many Grammatical Errors		.167			**.779**	
F6 Content richness						
16 Informative Speeches	.194					**.729**

* The numbers in bold indicate that the value of each item loaded on a specific factor is greater than .30.
ø The numbers in bold and italics indicate that the items loaded on more than one factor and their loading values are usually close to or greater than .30.
∞All loading values lower than .1 was omitted in all similar tables in this study.
All of the above representations apply to all similar tables in this study.

The reliability of each factor was assessed by means of *Cronbach's* alpha coefficients. Alpha coefficients provide a measure of internal consistency and an estimate of factor reliability. In Table 2, the alpha values of Factors 1, 2, and 3 in each factor solution were relatively high (around .8). The rest of the alpha values were relatively low. The low inconsistency of these factors explained the instability of teacher responses, while simultaneously demonstrating the complicated nature of constructing rating criteria for oral assessment. In short, based on the scree plot, the eigenvalue-one rule, and the reliability of each factor demonstrated *through Cronbach's* alpha coefficient, the number of factors retained was six.

Table 3 illustrates the rotated teacher factor pattern matrix. Items placed under a particular factor meant these items loaded on this factor. I then named the factor based on the common features shared by these items. As Table 3 shows, the six factors along which native Chinese evaluated the learner were as follows: communication appropriateness (F1), Chinese cognitive patterns (F2), communicative clarity (F3), listening comprehension (F4), language accuracy (F5), and content richness (F6).

The communication appropriateness or social acceptability factor mainly included five items with all factor loading values greater than .42. Item 11, an item dealing with content richness, double loaded on F1 and F6 with low loading values on both factors. In light of its similarity (in terms of content and meaning) to Item 16 within F6, I ascribed Item 11 to F6. Both items deal with content richness. Similarly, Item 13 double loaded on F2 and F5, but was ascribed to F5 considering meaning consistency. For example, Item 13, Wrong Word Usages and Item 7, Many Grammatical Errors shared the same meaning regarding accuracy. Therefore, when they loaded on one factor, they were ascribed to one factor (F5). However, the meaning of Item 13 was less consistent or not consistent at all with Item 4 (Significant Differences in Thought). Even if Item 13 double loaded with Item 4 on one factor, they were not regarded as one factor. Chinese cognitive patterns (F2) were made up of three items, each of which had a loading value greater than .62 and reflected Chinese

cognitive thinking patterns. F3 dealt with clarity from the perspective of communication. The three items with the highest loading values on F3 indicated the independence of positive responses with respect to clarity of communication as a rating criterion for oral assessment. Items dealing with listening comprehension, which consisted of two items with loading values greater than .59, loaded highly on F4.

3.2 EFA of Non-Teachers' Responses

The scree plot of the non-teacher data shows a break between F4 and F5, which suggests that four factors could be retained for this data. The eigenvalue-one rule also suggested that four factors could be retained, because Factors 1 through 4 had eigenvalues greater than one, with the eigenvalues of Factors 5 and 6 close to the cut-off point. To avoid underfactoring or overfactoring, EFA was conducted using four, five, and six factors respectively.

Table 4 shows the four- through six-factor solutions and reliabilities. The difference between the five-factor solution and the six-factor solution was that F4 (Items 7, 13) and F5 (Items 15, 9) in the six-factor solution were merged into F2 (Items 13, 7, 15, 9) in the five-factor solution. The difference between the four-factor solution and the five-factor solution was that F1 (Items 17, 5, 10, 12, 6) and F5 (Items 11, 16) in the five-factor solution were merged into F1 (Items 16, 17, 12, 10, 5, 11, 6) in the four-factor solution. Structure stability and consistency can be seen by comparing different factor solutions.

I retained the six-factor solution as the rating criterion pattern in this study. The reasons were as follows. First, the factors in the six-factor solution formed meaningful clusters. Combining any of the two factors in this solution, such as the combination of F4 and F5 into one factor in the four- and the five-factor solutions and the combination of F6 and F1 into one factor in the five-factor solution, made it more difficult to pinpoint a common meaning for the combined factors. Second, if the six-factor solution was chosen, then the pattern in the non-teacher data was consistent with the six-factor solution pattern in the teacher data. Accordingly, it was relatively easier to compare the two patterns.

Table 4: Factor Solutions of Non-Teacher Responses

Factors	# of Items	Items	Alpha
The four-factor solution			
#1 [o]	7	16[∞], 17, 12, 10, 5, 11, 6	.791
#2	5	13, 7, 15, 9	.783
#3	3	1, 3, 2	.740
#4	2	4, 8, 14*	.786
The five-factor solution			
#1	7	17, 5, 10, 12, 6	.754
#2	3	13, 7, 15, 9	.783
#3	3	3, 1, 2	.740
#4	2	8, 4, 14	.786
#5	2	11, 16	.683
The six-factor solution			
#1	5	17, 5, 10, 12, 6	.754
#2	3	3, 1, 2	.740
#3	3	8, 4, 14	.786
#4	2	7, 13	.708
#5	2	15, 9	.779
#6	2	11, 16	.683

[o] Each number in this column represents a factor. For example, #1 under the four-factor solution means Factor 1 in the four-factor solution.

[∞] Each number represents an item. For example, 16 represents Item 16. The description of each item is listed Table 1.

[*] Items within each factor are listed in the order of the values of factor loadings, from the highest factor loading value to the lowest one. For example, # 4 under the four-factor solution lists Item 4 before Item 8, because the loading value for Item 4 is higher than that of Item 8.

Additionally, the six-factor solution reflected the structure consistency and stability based on the two sets of data. Therefore, the numbers of factors in both data were retained as six.

As can be seen in Table 4, the alpha values of the factors were relatively high. The factor containing Items 11 and 16 had a value of .683 and

the factor containing Items 7 and 13 had a value of .708. The rest of the alpha values were around or above .75. According to Peter (1979), .60 is an acceptable level of reliability in the early stages of basic research for a proposed construct measurement. The relatively high values in this data (close to or greater than .70) demonstrated consistency and high reliability of those factors for non-teachers. In sum, based on the scree plot, the eigenvalue-one rule, and the reliability of each factor demonstrated *through Cronbach's* alpha coefficients, the number of factors retained was six.

Table 5 illustrates the rotated pattern matrix of non-teachers' responses. Items were also placed under each factor according to their factor loading values, which was the same as those in the teacher factor pattern and other rotated patterns in this chapter. The six factors along which non-teachers evaluated the speaker were: communication appropriateness (F1), communicative clarity (F2), Chinese cognitive patterns (F3), language accuracy (F4), listening comprehension (F5), and content richness (F6).

As for the factor of the communication appropriateness or social acceptability (F1), six items loaded on it: Items 17, 16, 5, 10, 12, 6 with the loading values higher than .4. Item 16 had a quite high loading value on F1 with .627, but it also loaded on F6 with a loading value of .410. I ascribed it to F6 instead of F1 in light of meaning consistency. Therefore, Item 16 together with Item 11 formed F6 of content richness. In the teacher data, F6 also consisted of the two items, but with Item 11 double loaded.

Table 5: Rotated Pattern Matrix of Non-Teacher Responses

Factors and Items	F1*	F2	F3	F4	F5	F6
F1 Communication appropriateness						
17 Appropriate Tone and Intonation	**.661**		.217		-.117	
16 Informative Speeches	***.627***	-.123	-.169		.185	***.410***
5 Communication Strategy	**.612**					
10 A Person Others Like to Deal with	**.534**			.219	-.243	
12 Appropriateness in Choosing Words and Structures	**.407**		.112	-.113	-.267	.202
2 High Proficiency	***.403***	***.309***	-.172			
F2 Communicative clarity						
3 Clear Expressions		**.871**				
1 Fluent Speech	-.119	**.832**				.274
F3 Chinese cognitive patterns						
8 Does Not Know Chinese Ways of Thinking	.161		**.887**			-.141
4 Significant Differences in Thought	-.220		**.600**			
14 Wrong Perceptions of Chinese Thinking	.101		**.496**	.160	.250	
F4 Language accuracy						
13 Wrong Word Usages				**.824**		.155
7 Many Grammatical Errors				**.693**		
F5 Listening comprehension						
15 Poor Listening					**.662**	
9 Does Not Understand the Interviewer				.138	**.465**	
6 Questions Easy For Him/her to Answer	***.226***	-.117			***-.326***	
F6 Content richness						
11 Rich Content		.280		.123	-.234	**.602**

*The abbreviation representations in this table can be found in Table 3.

Regarding Item 6, its loading value on F1was as low as .226, but I ascribed it to F1 because its loading values on other factors were also low. The highest loading of Item 6 was on the factor of

comprehension (F5), with a value of -.326, but the meaning was not consistent with the other items within F5. Based on the aforementioned considerations, I concluded that the items within F1 in the non-teacher data were the same as those within F1 in the teacher data.

The factor of communicative clarity (F2) was similar to that in the teacher data. It was made up of Items 3, 1, and 2. Item 2 double loaded on F1 and F2 and the meanings are consistent on both factors. The study ascribed it to F2, rather than F1 because the three items always loaded together, not only in the reported data, but also in other teacher and non-teacher response data, as reported in Chen (2011). However, the meanings of the three items were not consistent, I named the factor following the item with highest loading. Regarding F3, F4, F5, and F6, the items within each of the factors were the same across the teachers' and non-teachers' response data.

3.3 *Comparisons of Factor Structures of Teacher and Non-Teacher Responses*

Table 6: Comparisons between Teacher and Non-Teacher Rotated Factor Pattern Matrices

The Pattern in the CT Data		The Pattern in the NT Data
F1 (5, 10, 12, 6, 17)	≈	F1 (17, 5, 10, 12, 6)
F2 (8, 14, 4)	≈	F3 (8, 4, 14)
F3 (1, 3, 2)	≈	F2 (1, 3, 2)
F4 (9, 15)	≈	F5 (15, 9)
F5 (7, 13)	≈	F4 (7, 13)
F6 (16, 11)	≈	F6 (16, 11)

Note: NT represents Non-Teacher and CT represents Chinese Teacher.

As summarized in Table 6, the two groups responded similarly. If the six-factor solution was taken, the factor structure and the contained items were similar. F1 and F6 in the teacher group corresponded to those in the non-teacher group. F2 in the teacher group corresponded to F3 in the non-teacher group. In the same way, F3, F4, and F5 in the teacher group corresponded to F2, F5 and F4

respectively in the non-teacher group.

Pearson r

A Pearson product-moment correlation coefficient (Pearson *r*) can be used to compare two variables in terms of the magnitude and direction of association between them. Correlating the loadings on each factor in the teacher group with the loadings on the corresponding factor in the non-teacher group, a correlation matrix was obtained, as shown in Table 7. The correlations between each of the corresponding factors were .752, .874, .803, .726, .789, .639 respectively. The high correlation values indicated similarities between the two factor structures.

Table 7: Pearson R Correlation Matrix of **Teacher vs. Non-Teacher** Responses

Factors	NT_F	NT_F	NT_F	NT_F	NT_F	NT_F
CT_F1 Communication appropriateness	.752**	-.322	-.232	-.523*	-.218	.124
CT_F2 Chinese cognitive patterns	-.436	.803**	-.365	.089	.180	-.108
CT_F3 Communicative clarity	-.228	-.320	.874**	-.062	-.266	.071
CT_F4 Listening comprehension	-.397	-.088	-.020	.789**	.183	-.262
CT_F5 Language accuracy	-.248	-.031	-.150	-.079	.726**	-.201
CT_F6 Content richness	.304	-.297	-.222	-.056	-.098	.639**

Note:
** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).
NT represents Non-Teacher and CT represents Chinese Teacher.

Alpha Coefficients

The reliability of each of the six factors was assessed by means of Cronbach's alpha, as shown in Table 8.

Table 8: Reliabilities of Teacher, Non-Teacher, and **Teacher/Non-Teacher** Responses

Factors	# of Items	Items	CT Alpha	NT Alpha	CT/NT Alpha
The six-factor solution					
#1	5	5, 10, 17, 12, 6	0.783	0.754	0.723
#2	3	8, 14, 4	.809	.740	.779
#3	3	1, 2, 3	.809	.786	.798
#4	2	9, 15	.584	.708	.653
#5	2	7, 13	.672	.779	.730
#6	2	16, 11	.662	.683	.672

Note: NT represents Non-Teacher and CT represents Chinese Teacher.

These coefficients ranged from .584 to .809 in the teacher data, and from .683 to .785 in the teacher data. Alpha coefficients were also consistently high when the teacher and non-teacher data were combined, ranging from .653 to .798, which also suggests that **t**eachers' and non-teachers' **r**esponded to the speaker in a similar manner.

Mean Scores

Another comparison was conducted across the two groups by using the mean factor scores. Table 9 indicates that the mean factor scores of the teacher data were similar to those of the non-teacher data for all six factors.

Table 9: Means and Standard Deviations of All Factors

	Chinese Teacher			Non-Teacher		
	n	mean	S.D.	n	mean	S.D.
NT_F1 (CT_F1)	167	2.5186	.55633	176	2.4489	.49048
NT_F2 (CT_F3)	167	2.2595	.60165	176	2.3769	.63346
NT_F3 (CT_F2)	167	3.2066	.76534	176	3.1506	.65142
NT_F4 (CT_F5)	167	3.4611	.68546	176	3.3977	.71777
NT_F5 (CT_F4)	167	3.7784	.58238	176	3.7301	.63664
NT_F6 (CT_F6)	167	2.5749	.66002	176	2.5710	.70249

Note: NT represents Non-Teacher and CT represents Chinese Teacher.

A one-factor multivariate analysis of variance (MANOVA), summarized in Table 10, shows that the main effects of group were not significant (p=.142). In other words, teacher and non-teacher perceptions of the speaker's proficiency level did not significantly differ. Univariate analyses of variance were also performed on the data for each dependent variable (i.e., factor). A summary of these ANOVAs is given in Table 11. Results showed that there were no significant differences in ratings between each pair of the six factors.

Table 10: Summary of Multivariate Analysis of Variance (Wilks' Lambda)

Source	df	F	P
A (Group)	6, 336	1.616	.142

Table 11: Summary of Univariate Analysis of Variance of Each Dependent Factor

Source	df	SS	F	P
		NT_F1 (CT_F1) Communication appropriateness		
A (Group)	1	.416	1.519	.219
S/A	341	93.477		
Total	342	93.894		
		NT_F2 (CT_F3) Communicative clarity		

A (Group)	1	1.181	3.091	.080
S/A	341	130.311		
Total	342	131.492		
		NT_F3 (CT_F2) Chinese cognitive patterns		
A (Group)	1	.269	.535	.465
S/A	341	171.494		
Total	342	171.763		
		NT_F4 (CT_F5) Language accuracy		
A (Group)	1	.344	.697	.404
S/A	341	168.156		
Total	342	168.500		
		NT_F5 (CT_F4) Listening comprehension		
A (Group)	1	.200	.536	.464
S/A	341	127.233		
Total	342	127.433		
		NT_F6 (CT_F6) Content richness		
A (Group)	1	.001	.003	.959
S/A	341	158.677		
Total	342	158.678		

Note:
1. NT represents Non-Teachers and CT represents Chinese Teachers.
2. F1represents Factor 1 and so on.

4. Discussion

4.1 Main Findings

This study provides strong evidence that there is no significant difference between teachers and non-teachers in their rating criterion patterns for the speech performance of a foreign learner of Chinese. As shown in Table 6, teachers and non-teachers exhibited highly similar response patterns, in that the factors and the items contained in each factor were exactly the same. Although the order of items in each factor was not the same, and the order of importance of each factor accounting for the whole variance was different, the patterns of similarities are apparent for both teachers' and non-teachers' data. Correlational analyses also pointed to the similarities between the factors by showing high Pearson coefficients between each factor for

teachers and the corresponding factor for non-teachers (Table 7). Alpha coefficients were consistently high when the teachers' and non-teachers' data were combined, which suggested that teachers and non-teachers perceived the speaker's oral performances in a similar manner (Table 8).

A MANOVA showed no statistically significant difference between teachers' and non-teachers' responses overall (Table 10). ANOVAs for each factor within these response patterns demonstrated that there was no significant difference in teachers' and non-teachers' perceptions of the six factors (Table 11). All in all, analyses of teachers and non-teachers responses showed strong similarities between the responses of the two groups, and no significant differences between their rating criterion patterns. I believe that these similarities between teachers' and non-teachers' reflect shared Chinese cultural values, perspectives, and practices.

The data showed strong statistical evidence for similarities between teacher and non-teacher response. But this does not mean the groups were literally identical. For example, teachers' and non-teachers' ratings are all different for all rating criteria in Table 9. However, statistically, these differences are minor and can be ignored.

The purpose of this study was to test whether a teaching background influences native speakers' ratings of oral performances. The hypothesis was that it would not, based on the assumption that native speakers of Chinese are strongly influenced by shared cultural values, perspectives, and practices. and these cultural influences will dilute the effect of teaching experiences on perceptions of rating performances. The results provided evidence in support of this hypothesis.

4.2 Comparison of This Study to Other Studies

This study presents evidence for the absence of differences between teachers' and non-teachers' rating criterion patterns. This result is consistent with Hadden's work (1983, 1991). These studies all used similar research methods and found comparable patterns of similarities. Shohamy et al. (1992) compared the inter-rater reliability of four groups of raters: ESL teachers who receive training; ESL teachers who do not receive training; non-ESL teachers who receive training; and non-ESL teachers who do not receive training. These

researchers also found that teaching background does not make a difference.

The current study did not find significant differences between teacher and non-teacher raters for each pair of rating criteria. In contrast, Okamura (1995) concluded that teachers are more critical than non-teachers on most criteria. Hadden's results showed that teachers judge students' oral performance more critically than non-teachers do with respect to linguistic ability, but not on other criteria. Galloway (1977; 1980) concluded that teachers are more critical only with respect to pronunciation and rate of speech.

The reason for the differences among these findings might be related to differences in methodology, in the languages and cultures under study, in the target of evaluation (oral or written), and so on. For example, the focus of the current study on Chinese natives' perceptions may account for differences from studies in which the focus was on raters from western backgrounds. A key consideration is that prior research has tended to rely on small sample sizes. For example, in Hadden (1983, 1991) the entire sample consists of 25 teachers and 32 non-teachers. The sample size of the current study is 343. Having a large sample enabled the use of Exploratory Factor Analysis (EFA). Zhao (2009) lists the different minimum subject sizes necessary to perform EFA. For example, Hatcher (1994) recommends a sample of either 5 times the number of variables, or 100. Hutcheson and Sofroniou (1999) recommends at least 150 to 300 cases. Garson (2008) recommends at least 300 cases. Larger samples are evidently needed in order to use advanced statistical analyses to evaluate the reliability of the rating criterion structure.

One further possible explanation for the divergence among studies pertains to instructions given to raters. In the present study, raters do not receive pre-established rating criteria when evaluating the speech sample. Other studies, such as Okamura (1995), compared teachers and non-teachers rating on given rating criteria (grammar, fluency, appropriateness, vocabulary, comprehensibility, and pronunciation).

4.3 Implications of the Study

The study differs from other studies by presenting evidence for similarities among teachers' and non-teachers' ratings. It is critical to note that in this study, teachers' and non-teachers' rating behaviors were based on relatively immediate perceptions rather than on a careful and sustained analysis of oral performances and a correspondingly careful and sustained consideration of the rating criteria. The perceptions in question are influenced by Chinese culture, particularly cultural perspectives (i.e., attitudes, beliefs, or values shared by many people in a culture). The similarity between teacher and non-teacher responses in this study reflects the fact that cultural factors are important and dominant components in perceiving language-related proficiency. Because these perceptions are somewhat ingrained, they are stable and somewhat resistant to change.

Rating performance is subjective in nature. Naïve native speakers' rating performance is even more unpredictable and inconsistent. However, the individually different rating behaviors share the same culture values, practices, and perspectives, thereby demonstrating consistency at a macro level. In light of the results of this study, I argue for the importance of studying native speakers' perceptions, especially, the culturally inbuilt values, practices, and perspectives, all of which can inform experts in Chinese as a foreign language assessment.

Experts in language assessment should make studying the underlying cultural assumptions and attitudes informing the responses of teachers and non-teachers a top priority. These experts should be equipped in language assessment to judge the effectiveness of a non-native speaker's communicative effectiveness. However, they should also need to be aware of how naïve native speakers perceive learners' performance outside the classroom.

4.4 Limitations and Future Studies

A famous dictum in philosophy holds that absence of evidence does not equal evidence of absence. The purpose of this study was to present evidence for the absence of differences between the teacher and non-teacher groups. This strategy carries the risk that there are

differences between the two groups that were not revealed in my data. I addressed this issue by means of multiple approaches to data analysis.

Another potential limitation is that the teacher sample in the present study does not represent all possible teachers of Chinese. The teacher sample was mainly composed of Chinese teachers in the U.S. sponsored by the office of the Chinese Language Council and the Chinese teachers who are on the Chinese Language Association mailing list. Many experienced Chinese teachers in the U.S. and many teachers in mainland China are excluded from this study. Nonetheless, the teachers that comprised the teacher sample in this study are reflective of a broad spectrum of Chinese teachers.

Future studies might also address this issue by directly contrasting teacher and non-teacher ratings with those of experts in Chinese language assessment of oral performance. Another approach in future research could diversify the teacher group by adding many experienced Chinese teachers in the U.S., mainland China, and other oversea areas. Then the evidence that the two groups are comparable would be stronger.

Finally, a limitation is reflected in the oral performance sample. This paper only reported native speakers' responses to one speech sample. Ideally, more samples, and a longer duration of samples, would be preferable. However, the focus of this paper is to test native speakers' perceptions, which requires limited sample length. In addition, the large sample of respondents (343) limited the number of speech samples that could be used. However, native speakers' responses to the speech sample used in this study are similar to their responses to two other samples, as reported in Chen (2011). The similarity across findings suggests that representativeness of the speech sample used in the current study. Future studies might address this issue through analysis of transcripts of oral performances, which might provide more information about sample representativeness.

References

Barnwell, D. (1989). 'Naive' native speakers and judgments of oral proficiency in Spanish. *Language Testing, 6*(2), 152–163.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12(1),* 1-15.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing,* 12, 16–33.

Chen, G. (2011). *Developing a culture-based rating criterion model for assessing oral performances in teaching Chinese as a foreign language.* PhD thesis. Ohio State University.

Child, D. (1990). *The essentials of factor analysis* (2nd ed.). London: Cassel Educational Limited.

Elder, C. (1993). How do subject specialists construe classroom language proficiency? *Language Testing, 10(3),* 235-254.

Fayer, J. M. & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning,* 37, 313–326.

Galloway, V. B. (1977). *Analysis of evaluations of the oral communicative competence in Spanish of University of South Carolina students.* Dissertation Abstracts International 38, 7255A (University Microfilm 78-07, 900).

----. (1980). Perceptions of the communication efforts of American students of Spanish. *Modern Language Journal, 64,* 428-433.

Garson, D. G. (2008). *Factor Analysis: Statnotes.* Retrieved March 22, 2008, from North Carolina State University Public Administration Program.

Hadden, B. L. (1983). A comparison of teacher and non-teacher perceptions of the second language communication of advanced ESL students who are native speakers of Chinese. PhD thesis. The Ohio State University.

----. (1991). Teacher and nonteacher perceptions of second language communication. Language Learning, 41, 1-24.

Hatcher, L. (1994). *A Step-by-Step Approach to Using the SAS® System for Factor Analysis and Structural Equation Modeling.* Cary, NC: SAS Institute, Inc.

Hutcheson, G. D, & Sofroniou, N. (1999). *The multivariate social scientist: Introductory statistics using generalized linear models*. Thousand Oaks, CA: Sage Publications.

Jacoby, S., & McNamara, T. F. (1999). Locating competence. *English for Specific Purpose, 18(3),* 213–241.

Kim, Y. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*, 187-217.

Okamura, A. (1995). Teachers' and non teachers' perception of elementary learners' spoken Japanese. *The Modern Language Journal, 79*, 29-40.

Palmer, L. A. (1973). Preliminary report on a study of the linguistic correlation of raters' subjective judgments of non-native English speech. In Roger W. Shuy and Ralphw Fasold (Eds.), *Language attitudes: Current trends and prospects*. Washington, D.C.: Georgetown University Press.

Peter, P. J. (1979). Reliability: A review of psychometric basics and recent marketing practices. *Journal of Marketing Research*, XVI (Feb.), 6-17.

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, *18*, 303–325.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effects of raters' background and training on the reliability of direct writing tests. *Modern Language Journal, 76 (1),* 27-33.

Walker, G. (2000). Performed culture: Learning to participate in another culture. In Richard Lambert and Elana Shohamy (Eds.), *Language policy and pedagogy* (pp. 221-236). Philadelphia: John Benjamins Publishing Company.

Walker, G., & Noda, M. (2000). Remembering the future: Compiling knowledge of another culture. In Diane Birckbichler and Robert Terry (Eds.), *Reflecting on the past to shape the future* (ACTFL Foreign Language Education Series) (pp. 187-212). Lincolnwood, IL: National Textbook Co.

Wiggins, G. (1989). A true test: toward more authentic and equitable assessment. *Phi Delta Kappan*, *70(9),* 703-713.

Zhao, N. (2009). The minimum sample size in factor analysis, available at

http://www.encorewiki.org/display/~nzhao/The+Minimum+Sample+Size+in+Factor+Analysis, accessed on September 8, 2013.