Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 6, No. 2, October 2016
ISSN 2476-5880

Developing a Model of Analytic Rating Scales to Assess College Students' L2 Chinese Oral Performance

Guangyan Chen[1]

Received: 8 May 2016 Accepted: 21 August 2016

Abstract

This study develops a model of analytic rating scales to assess L2 Chinese oral performance. It uses Exploratory Factor Analysis (EFA) to identify a model and employs Confirmative Factor Analysis (CFA) in a separate dataset to test the degree of model fit. The researcher videotaped ten speeches and ACTFL professional raters assessed the oral performances in these samples. The researcher then selected three samples (Samples 1, 2, and 3) to represent the proficiency levels of Novice High, Intermediate High, and Advanced Low. Then, the researcher developed 20 rating items by interviewing ten experienced L2 Chinese teachers and running an EFA. The 20 items were descriptors that Chinese teachers used to assess oral performance in two studies: Study 1 and Study 2. To complete Study 1, the researcher recruited 45 teachers to assess Sample 1 using the 20 items, 62 teachers rated Sample 2, and 49 teachers rated Sample 3. In Study 2, 104 teachers assessed all three samples. The EFA indicated a four-factor model of analytic rating scales: "fluency," "conceptual understanding," "communication clarity," and "communication appropriateness." In this model, the correlations between these analytic rating scales were relatively high and teachers weighted "fluency" as most important. Together the four scales explained 65.5% of teachers' holistic judgments of oral performance. The CFA did not show a strong model fit to the data, but the fit was acceptable. This model advances our understanding of the relationship between analytic rating scales and holistic ratings in the context of L2 Chinese. These findings give Chinese teachers with which a reference to assess U.S. college students' L2 Chinese oral performance.

Keywords: *Analytic rating scales, Factor analysis, Language assessment, L2 Chinese, Oral performance*

1. Introduction

In the field of performance-based language assessment, many teachers and assessment professionals adopt analytic rating scales, holistic rating scales or a combination of both to assess oral performance. A holistic scale evaluates the overall quality of language performance. It offers advantages, such as easy score reporting and score efficiency (Fulcher, 2010; Xi, 2007). However, a holistic scale can be problematic, because it may not provide much information about the link between the descriptors and the language performance (Fulcher, 2003; 2010).

Analytic rating scales provide multiple scores for language performance, which indicate the multi-componential nature of language competence. These scores provide useful diagnostic

[1]*The Department of Modern Language Studies, Texas Christian University, USA. E-mail: g.chen@tcu.edu*

information for test users (Fulcher, 2010). Analytic rating scales also allow for the possibility of generating a single composite holistic score. One can obtain the composite score by averaging analytic scores or weighting these scores differentially (Kondo-Brown, 2002; Sawaki, 2007; Weigle, 1998; Xi, 2007). Scholars have well documented the advantages of analytic over holistic rating scales (Bachman et al., 1995; Bachman & Savignon, 1986; Kondo-Brown, 2002; Sawaki, 2007; Xi, 2007). A common advantage, as Bachman and Savignon (1986) argued, is that language ability is multi-componential, so analytic ratings are better than a holistic rating in reflecting language ability. However, it is difficult to operationally define language ability, because researchers have different understandings of it and what it entails (Bachman, 1990; Canale, 1983; Canale & Swain, 1980; Chomsky, 1965; Hymes, 1972; Saussure, 1959; Walker, 2000; Young, 1999). Therefore, the relationship between analytic ratings and corresponding holistic ratings is a lively research topic in the field of foreign language and/or second language testing. The current study contributes to this line of research by examining this relationship in the context of L2 Chinese oral assessment**.**

A rating scale is a scoring guide used to assess performance against a set of criteria. According to Hudson's (2005) definition, a scale1) reflects a continuum of performance quality; 2) identifies the significant traits or dimensions being assessed; 3) provides key performance criteria for each level of scoring in "descriptors," which reflect the extent to which the key performance requirements have been demonstrated (Hudson, 2005: 208). Analytic rating scales, in this study, reflect the three aspects of Hudson's definition of scale but focus on the third: providing key criteria for assessing oral performance. In essence, the term "analytic rating scales" equals to the term "rating criteria" throughout the study. Some scholars in linguistics (e.g., Iwashita et al., 2008; Jin & Mak, 2013) adopt the term "performance features," or "features," which focuses on the second aspect of Hudson's definition. Again, the term "performance features," or "features," refers to analytic rating scales in my study.

Many previous studies (Jin & Mak, 2013; Plough et al., 2010; Sawaki, 2007; Xi, 2007) adopt, rather than develop, analytic rating scales and examine the relationships between analytic rating scales and their corresponding holistic ratings in assessing oral performance. The above studies address questions concerning the relationships between their adopted analytic rating scales and the corresponding holistic ratings and the relationships among these scales. Yet, few studies examine such questions as how many criteria in these scales are optimal to assess oral performance, which criteria are used to assess oral performance, and the degree to which these criteria explain raters' holistic judgments of oral performance. This study narrows these research gaps.

In the context of L2 Chinese, I locate four sets of documents that involve analytic rating scales for assessing speaking:

1. *ACTFL Chinese Proficiency Guidelines* (ACTFL, 1987)
2. *Chinese Language Proficiency Scales for Speakers of Other Languages* (The Office of Chinese Language Council International, 2007)
3. *International Curriculum for Chinese Language Education* (The Office of Chinese Language Council International, 2008)
4. *Spoken Chinese Proficiency Grading Standards and Testing Guidelines* (Ministry of Education & State Language Commission, the People's Republic of China, 2011)

The four rating scales listed above, including proficiency levels and descriptors for each level, are similar to those scales in other proficiency guidelines or curricula, such as the Foreign Service Institute (FSI) scales, the ACTFL proficiency guidelines, and the Canadian Language Benchmarks. All these rating scales are holistic in nature (Fulcher, 1996) because the weights of each criterion in these scales and the relationships between these criteria are not determined. This context in L2 Chinese oral assessment outlines the need for this article.

2. Literature review

2.1. Number of analytic rating scales

When developing analytic rating scales for assessing language performance, developers need to decide how many criteria within these scales to consider. According to the suggestions from The Common European Framework of Reference for Languages (Council of Europe, 2001) and Luoma (2004), four or five categories begin to cause a cognitive load for raters and seven categories are considered a psychological upper limit. Five or six categories may be close to maximum. However, previous studies have not provided empirical evidence to support the determination of optimal number of criteria within rating scales.

2.2. Content of analytic rating scales (Or which analytic rating scales are used to assess oral performance?)

In the field of oral performance assessment, analytic rating scales vary according to the purpose of tests. Researchers disagree on which analytic rating scales that the language testers and teachers should use to assess daily conversation(Adams, 1980; Hadden, 1991; Higgs & Clifford, 1982; Jin & Mak, 2013; Sato, 2012; Wang, 2002).For example, Hadden (1991) performed Exploratory Factor Analyses (EFAs) to compare ESL (English as a Second Language) and non-ESL teachers' perceptions of eight ESL learners' speaking performances. Hadden discovered that the ESL and non-ESL teachers relied on a similar rating model consisting of the following analytic rating scales: comprehensibility, social acceptability, linguistic ability, personality, and body language. In order to assess speech samples from the FSI oral interview, Adams (1980) investigated the contribution of five analytic rating scales (accent, comprehension, vocabulary, fluency, and grammar) to a holistic speaking score. Higgs and Clifford (1982) proposed a five-scale model (vocabulary, grammar, pronunciation, fluency, and sociolinguistics) by describing rater perceptions. In the context of L2 Chinese, the studies of Jin and Mak (2013) and Wang (2002) examined the relationship between the performance features—pronunciation, fluency, vocabulary, and grammar—and holistic ratings. Each of the two studies focused on slightly different performance features and used different methods, but both confirmed a general connection between these features and holistic ratings. However, these features were limited to linguistic components. Other components indicating communicative ability, such as the content component, were rarely explored (Sato, 2012).Sato found that the content component was an important scale for assessing speech in addition to the linguistic components, such as grammatical accuracy, fluency, vocabulary range, and pronunciation.

2.3. Weights of analytic rating scales

When testers attempt to arrive at a single composite score based on componential scores, the relative contributions of various analytic rating scales are important. However, various language competence models (Bachman, 1990; Canale, 1983; Canale & Swain, 1980; Chomsky, 1965; Hymes, 1972; Saussure, 1959; Walker, 2000; Young, 1999) differ in their implications for weighting various analytic rating scales. Even though there are studies on weights of analytic rating scales in ESL or other foreign languages (e.g., Bachman & Palmer, 1981; Iwashita et al., 2008; Plough et al., 2010; Sato 2012; Sawaki, 2007), there has been little empirical research in L2 Chinese oral assessment. Below are the studies examining the weights of analytic rating scales. Plough et al. (2010) used Stepwise Logistic Regression and concluded that pronunciation and listening comprehension were most important in predicting the speaking proficiency of prospective graduate student instructors. Sato's (2012) study revealed that the content component made a substantive contribution to holistic judgments. Sawaki (2007) used the approaches of Confirmative Factor Analysis (CFA) and multivariate Generalizability (G) theory and identified the scale of grammar as the largest contribution to the composite score.

Some studies demonstrate that the weights of analytic rating scales vary across different proficiency levels (Adams, 1980; Higgs & Clifford, 1982; De Jong & Van Ginkel; 1992; Iwashita et al., 2008).Adams (1980) found that accent and fluency were not significant determinants of holistic scores at lower levels of proficiency, while vocabulary and grammar played a significant role. However, as proficiency levels increased, other factors became more important. Similarly, Higgs and Clifford (1982) also pointed out that vocabulary and grammar were important across all proficiency levels. However, at lower levels, teachers perceived vocabulary and pronunciation as two more important scales. At higher levels, the sociolinguistic scale was relatively less important than the other four scales. De Jong and Van Ginkel (1992) discovered that pronunciation was most strongly predictive of holistic ratings at lower levels of proficiency, whereas fluency became more predictive when proficiency levels went up. When developing rating scales for a new international test of English proficiency for academic purposes, Iwashita et al. (2008) investigated a global score and its relationship to detailed features (grammatical accuracy and complexity, vocabulary, pronunciation, and fluency) of the spoken language produced by test takers. They observed that each feature helped assess the overall levels of performance. The particular features of vocabulary and fluency had the strongest impact. All of the above studies in this subsection reached inconclusive results concerning the weights of individual analytic rating scales.

2.4. Correlations between analytic rating scales

Few studies examined the correlations between analytic rating scales (Sawaki, 2007; Xi, 2007). Sawaki (2007) investigated the construct validity of analytic scales in a speaking assessment. Sawaki used CFA and G theory in an analysis of 214 students' Spanish speaking performances to determine placement in a study abroad program. The results demonstrated strong correlations among the five analytic rating scales: pronunciation, vocabulary, cohesion, organization, and grammar. Xi (2007) explored empirically the utility of analytic scoring for TOEFL Academic

Speaking Test by performing G studies to investigate the dependability of the analytic scores, the distinctness of the analytic dimensions, and the variability of analytic score profiles. Xi (2007) also observed relatively high correlations among the analytic scores. My study continues along this line of research into L2 Chinese oral assessment.

Within this context, the current study extends previous research in a number of ways. First, I conducted EFA to develop analytic rating scales rather than to examine the relationships between adopted analytic rating scales and their corresponding holistic ratings. Second, I conducted an EFA to explore a model of analytic rating scales using one dataset and then performed a CFA to test the model fit using another dataset. This research method differs from previous studies, in that these studies either rely on EFA to explore rating models (Hadden, 1991) or use CFA, G theory, or other techniques to verify the construct validity of adopted analytic rating scales (Bachman & Palmer, 1981; Sawaki, 2007; Xi, 2007). Third, this examination of analytic rating scales used to assess L2 Chinese oral performance is significant because no previous studies have provided evidence concerning whether the scales developed for assessing ESL or other foreign languages can apply to L2 Chinese. It is well known that Chinese is a truly foreign language for U.S. learners. Chinese is a member of the Sino-Tibetan language family and completely unrelated to the Indo-European language family, a category to which English and most other European languages belong. The characteristics of the Chinese language and Chinese people's view of assessing speaking might differ from western languages and western assessors. Specifically, five main research questions guide this study:

1. How many analytic rating scales are retained?
2. What comprises the content of these analytic rating scales? Or which analytic rating scales are used to assess oral performance?
3. What are the correlations between these analytic rating scales?
4. To what extend do these analytic rating scales explain teachers' holistic judgments of oral performance?
5. What is the degree of model fit?

Among the above five research questions, the answers to the first four comprise the four aspects of the model of analytic rating scales. This study uses the answer to the fifth question to test the degree of model fit.

3. Method

3.1. Instruments

3.1.1. The three speech samples

I videotaped ten speakers during their OPI (Oral Proficiency Interview). The speakers were ten American students who were studying Chinese in different levels of language classes at a Mid-western university. Each video lasted about five-minutes and corresponded to each speaker. I sent the ten videos to three professionals, who had ACTFL OPI Rater Certification and assessed the ten videos according to the ACTFL proficiency guidelines (2009). Based on their

assessments, I randomly selected three samples to represent the levels of Advanced Low (Speech Sample 1), Intermediate High (Speech Sample 2), and Novice High (Speech Sample 3). Each sample covered seven topics**.** These topics included self-introduction, hometown description, family income, the impact of economic crisis, study abroad experiences, and comparisons between Chinese and American cultures and societies. The difficulty of topics ranged from simple personal questions to more difficult questions about social and cultural issues.

3.1.2. Development of rating items to assess speech samples

The term "rating item" in this study referred to descriptors, statements, or comments that a teacher used to assess oral performance, such as: "He responds to questions appropriately." I created the initial set of rating items by interviewing ten experienced L2 Chinese teachers. I based the interviews on a methodological question: "What relevant questions do you ask yourself when assessing a student's oral performance?" For example, one of teachers' responses to the question could be: "Does he/she respond to questions appropriately?" The corresponding item could be "He responds to questions appropriately." If the ten L2 Chinese teachers' responses showed that this item shared a similar meaning to others, I included it in the initial set; otherwise, I deleted it.

After this initial step, I analyzed the set of items through a pilot study, in which I recruited 42 Chinese students and scholars at a Mid-western school to assess Speech Sample 1. They used the initial set of 35 rating items to assess Speech Sample 1. I subsequently performed an EFA on their responses. After running the EFA, I deleted the rating items that did not cluster meaningfully with others. The 20 rating items shown in Table 1 were retained for further analysis in two subsequent studies: Study 1 and Study 2.

3.2. Participants: L2 Chinese teachers as evaluators

I sent the three speech samples and my questionnaires (the appendix) to L2 Chinese teachers for their responses and assessments. These teachers had at least one year of Chinese teaching experience in the U.S., either at the K-12 or the college level. They were all native speakers of Chinese. College Chinese teachers were the ideal participants, because the purpose of this study was to assess college-level student oral performance. I also looked for teachers who had many years of teaching experience. However, identifying an adequate pool of experienced L2 Chinese teachers at the college level was difficult. Therefore, I eventually expanded my participant base, including someK-12 Chinese teachers and the teachers with one year of teaching experience.

In Study 1, 45, 62, and 49 teachers assessed Speech Samples 1, 2, and 3 respectively, for a total of 156 responses. In Study 2, I sent all three speech samples and the corresponding questionnaires to a different set of L2 Chinese teachers for their assessments. I required the teachers to respond to all three speech samples and analyzed only complete responses. In total, 104 teachers responded to all three speech samples. The participants in Studies 1 and 2 were different, but both were L2 Chinese teachers at the K-12 and/or the college levels.

Table 1. The 20 items retained for analysis in Study 1 and Study 2

4*. His/her communication proficiency level is high ø.
1. ♯He/she has a good listening comprehension. 7. He/she can understand the questions. 12. He/she does not understand the questions being asked. 18. His/her listening comprehension is poor.
2. He/she has a good personality. 8. He/she employs effective communication strategies. 13. He/she is a person that others would like to deal with.
5. He/she expresses him/herself clearly. 20. I do not understand what he/she says.
6. His/her conceptual understanding is very different from that of a native speaker of Chinese. 11. He/she does know Chinese conceptual structure and verbal behaviors. 17. He/she perceives Chinese conceptual framework incorrectly.
10. He/she makes many grammatical errors. 16. He/she uses many words incorrectly. 15. He/she selects appropriate words and structures.
3. He/she speaks fluently. 9. These questions seem easy for him/her to answer. 14. The content of his/her answer is informative. 19. He/she delivers much information.

Notes. * The number before each item denotes the order that appeared in the questionnaire (the appendix). In the questionnaire, I scattered items that shared a similar meaning to decrease experimental errors.

ø The sentence is an English translation of a rating item that is written in Chinese in the appendix.

♯In Table 1, I grouped the items that shared a similar meaning in one cell. I hoped to see them clustering in the subsequent EFAs. Some of them, such as Items 7 and 12, related to the same content from both positive and negative perspectives.

3.3. Data analysis

In this study, the main analytical methods include EFA and CFA. I used EFA to explore a possible underlying factor structure (referred to here as the rating scale model) of an observable item set without imposing a preconceived structure on the outcome (Child, 1990). One important step of using observable items to predict hidden factors and a corresponding underlying structure is to determine the number of factors. The next step is to label these factors by analyzing the common features of rating items loaded on each factor. This study adopted the most frequently used and highly reliable methods to determine factor numbers: the Cattell scree plot and the eigenvalue-one rule. The scree plot graphs the eigenvalues (y-axis) of all the factors (x-axis) through listing them in decreasing order. The heuristic is to retain all the factors above (i.e., to the left of) the inflection point (i.e., the point where the curve starts to level off), because all the factors above the inflection point explain large proportions of variance. The factors after the inflection point explain a very small proportion of the variability and can be ignored. The eigenvalue-one rule is another hint of how many factors should be retained. The factors with Eigenvalues > 1 indicate their importance in interpreting large proportions of variance. Therefore, these factors can be retained. After the factor number is determined, the second step is to label each factor based on the common features of all items that clustered under that factor. When labeling factors, researchers commonly pay particular attention to those items with the highest loadings, because high loadings denote that those items have close association items with a hidden factor. I used one type of oblique rotation, promax rotation, in this study. I selected an oblique (rather than orthogonal) rotation because it assumes that factors correlate with each other. This approach provides a more realistic solution in the construction of rating scales, because rating scales are often correlated in the real world.

CFA is used to verify the factor structure of a set of observable items. CFA allows a researcher to test the hypothesis that a relationship exists between observable items and their underlying latent constructs. In other words, CFA is used to test the degree of fit between a proposed structural model and the emergent structure of the data. In this study, the model explored through EFA was subsequently evaluated by CFA on the basis of multiple criteria: (1) the appropriateness of the solutions, and (2) goodness of fit to the data. The criteria for evaluating goodness-of-fit are as follows:

- The ratio of Satorra-Bentler model chi-square to model degrees of freedom ($\chi^2_{S\text{-}B}/df$): There is no clear-cut rule about a cutoff point for this statistic, 5.0 or below usually is regarded as a good model fit.
- *Goodness of Fit Index (GFI)*: An absolute model fit index. A GFI of .90 or above implies an adequate model fit.
- *Comparative Fit Index (CFI)*: An incremental fit index, CFI assesses overall improvement of a proposed model over an independent model. A CFI of .90 or above indicates an adequate model fit.
- *Root Mean Square Error of Approximation (RMSEA)*: A RMSEA evaluates the extent to which the model approximates the data. A RMSEA of .05 or below is an indication of close fit, and a value of .08 or below as a signal of adequate fit (Browne & Cudeck, 1993). Usually, .1 or below denotes an acceptable value.

3.4. The assumptions of running the two studies

This study hinges on the following two assumptions: First, language competence is multidimensional and the corresponding rating should be based on analytic rating scales. Scholars have well documented this assumption (Bachman, 1990; Bachman & Palmer, 1996; Douglas & Selinker, 1992; 1993). Second, proficiency levels, which remain within the daily communication range, do not affect the model of analytic rating scales. Chen's study (2011) has provided empirical evidence to argue for this assumption. The ACTFL guidelines also support this assumption. One of the purposes of the guidelines are to measure college students' daily communication ability after they finish several years of foreign language learning. The guidelines use the same four scales across different proficiency levels: function, content, context, and accuracy.

4. Results

This section reports the EFA results for teachers' assessments of the three speech samples, followed by the CFA results used to test model fit. The EFA included four steps: 1) determining the number of factors (i.e., rating scales), namely, how many analytic rating scales I retained to assess oral performance; 2) naming each of these factors by analyzing the items clustered under that factor; 3) presenting the correlations between these factors; and 4) calculating the percentages of these factors to explain the overall judgment of these oral performances.

4.1. How many analytic rating scales are retained?

I initially used scree plot and the eigenvalue-one rule to retain the number of rating scales. The scree plot shows an inflection point between Factors 3 and 4, which suggests that Factors 1, 2, and 3could be temporarily retained for this data The reasons for retaining these factors were stated in the method section. The eigenvalue-one rule also suggests that these three factors could be retained, as Factors 1 through 3 had eigenvalues greater than one, denoting that Factors 1-3 explained most variance of the holistic rating. The eigenvalue of Factors 4 and 5were close to the inflection point.

To avoid underfactoring or overfactoring, I not only performed EFAs with three factors as suggested by the results of scree plot and the eigenvalue-one rule, but I also ran EFAs consisting of four and five factors. When I chose three, four, and five factors to run EFAs, respectively, I obtained three solutions: three-, four-, and five-factor solutions. Comparing the three-and four-factor solutions, I observed that the items clustered under one factor (F1[2]) in the three-factor solution did not share a consistent meaning. These items in the four-factor solution (Table 1) clustered under two factors, with one factor indicating "fluency" and the other

[2]F# represents Factor #, such as F1 representing Factor 1. This notation applies to the whole study.

implying "clarity." I, therefore, chose the four-factor solution. I also compared the four-and five-factor solutions and chose the four-factor solution for the following reason. The first four factors in the five-factor solution were similar to the four factors in the four-factor solution, yet the items under F5 in the five-factor solution did not share a consistent meaning with each other. In addition, their factor loadings were low. Based on these findings, I selected the four-factor solution for the next analysis.

The reliabilities of the four factors in the four-factor solution were assessed by means of Cronbach's alpha coefficient. Alpha coefficient provides a measure of internal item consistency and an estimate of factor reliability. The alpha values of Factors 1, 2, 3, and 4 were relatively high (.950, .858, .839, .710). The high reliability of these four factors further confirmed that four factors could be an optimal number retained for assessing oral performance.

4.2. Which analytic rating scales are used to assess oral performance?

Based on the scree plot, the eigenvalue-one rule, the comparisons of different factor solutions, and high factor reliability, the most detailed and meaningful clusters of items occurred when the number of factors retained was four. I labeled the four factors: fluency (F1), conceptual understanding (F2), communication clarity (F3), and communication appropriateness (F4) (see Table 2).

As shown in Table 2, eight items clustered under F1 "Fluency." I labeled F1 based on the items with the highest loadings: Items 14, 19, 3, 9, and 4. Items 14 (Rich Content) and 19 (Informative Speech) both related to content richness. Item 3 (Fluent Speech) implied the connotation of fluency. The description of Item 9 in the questionnaire was "These questions seem easy for him/her to answer," which could be categorized into "fluency," as "ease" or "automaticity" the characteristics that were usually regarded as one aspect of "fluency." Item 4 (High Proficiency) denoted an overall impression of proficiency. My decision of labeling F1"fluency" drew from Lennon (1990).Lennon (1990) provided a broad definition and a narrow definition of fluency. His broad definition states that fluency could be seen as overall speaking proficiency, whereas fluency in the narrow definition pertains to smoothness and ease of information delivery. Lennon's broad definition demonstrates that the connotation of fluency could be a very broad concept, including the overall impression of proficiency. In other words, Item 4 (High Proficiency) could imply "fluency." According to Lennon's narrow definition, Item 9 referred to ease of speaking, which was also associated with "fluency." In this study, I added one more dimension of fluency, the pace of speaking, because speaking more quickly may indicate the delivery of more content within a certain amount of time. I, therefore, ascribed "content richness" to the category of "fluency."

 ISSN 2476-5880

Table 2.Rotated pattern matrix of teachers' assessments

Factors and Items	1	2	3	4
F1 Fluency				
14 Rich Content	**.915**	-.025	-.194	-.042
3ø Fluent Speech	**.831**	-.007	.094	.019
19 Informative Speech	**.822**	-.041	-.052	.020
4 High Proficiency	**.787**	.178	.227	.077
9 Questions Easily Answered	**.766**	.097	.117	-.070
10 Frequent Grammatical Errors	***-.694****	***.300***	.186	.107
5 Clear Expression	**.628**	-.100	.053	.078
15 Appropriateness in Choosing Words and Structures	**.621**	.027	.250	.125
1 Good Listening	**.577**	-.048	.283	-.059
F2 Conceptual understanding				
6 Significant Differences in Conceptual Understanding	-.044	**.774**	.104	-.174
17 Wrong Perceptions of Chinese Conceptual Thoughts	.136	**.746**	-.151	-.031
11 Knowledge of Chinese Conceptual Thoughts	-.007	**.661**	-.194	-.120
16 Frequent Word Usage Errors	-.379	**.464**	-.084	.121
F3 Communication clarity				
18 Poor Listening	-.096	.094	**-.815**	.065
12 Not Understanding the Questions	-.309	.124	**-.575**	.201
7 Question Comprehension Ability	.271	.050	**.494**	.186
20 Utterances Not Understandable	-.132	.254	**-.439**	.145
F4 Communication appropriateness				
2 Good Personality	-.009	-.090	-.227	**.711**
13 A Person Others Like to Deal with	-.100	-.097	.041	**.683**
8 Communication Strategies	.305	.062	.200	**.533**

ø 3 represents Item 3 under Factor 1. This notation applies to the whole study.
* The number in bold and italics indicates that the item loaded on more than one factor. I assigned it to a factor considering its meaning consistency with other items under this factor.

F2 consisted of five items (Items 6, 17, 11, 16 and 10[3]). Items 10 and 16 dealt with language accuracy. Items 6, 17, and 11 related to people's conceptual understanding. I named F2 according to the shared meaning of the three items with highest loadings. Items 10 and 16 had relative low loadings with .464 and .300. Their shared meaning of language accuracy related to conceptual understanding. In this study, the terms "conceptual understanding," "conceptual thoughts," and "conceptual frameworks" all related to "conceptual structure" in Jackendoff's article (2002). According to Jackendoff (2002), conceptual structure is not a part of language per se, but part of thought. "Conceptual understanding" provides the locus of understanding linguistic utterances in context and incorporates pragmatic considerations and "world knowledge." It is cognitive structure that regulates linguistic allocation patterns and therefore directly relates to language accuracy. For this reason, I named F2 "conceptual understanding."

F3, consisting of Items 18, 12, 7, and 20, was labeled "communication clarity" based on the shared meaning of these items: Items 18 (Poor Listening), 12 (Not Understanding the Questions), and 7 (Questions Comprehension Ability), all of which related to listening comprehension, the ability to receive information. Item 20 (Utterances Not Understandable) was associated with comprehensibility of utterances. All four items implied communication clarity.

F4 consisted of Items 2 (Good Personality), 13 (A Person Others Like to Deal with), and 8 (Communication Strategies). It was labeled "communication appropriateness" because the three items all implied this concept.

4.3. What are the correlations between these analytic rating scales?

Table 3 shows the correlations between the identified factors. The overall correlations between each factor were quite high. In total, there were six correlations between the four factors. One correlation value of .759 occurred between "fluency" and "communication clarity." Other four correlations, such as those between F1 and F2 (-.599), F1 and F4 (.446), F2 and F3 (-.532), as well as F3 and F4 (.460), were also relatively high. One correlation between F2 and F4 (-.266) was low. The correlations between F1 and other factors were relatively high while the correlations between F4 and other factors were relatively low in the correlation matrix.

4.4. To what extend do these analytic rating scales explain teachers' holistic judgments of oral performance?

Table 4 illustrates the total variance regarding teachers' assessments of these speech samples. In short, four factors were extracted from the 20 items. The four factors explained approximately 65.5% of the total variance, which originally can be explained 100% from the 20 items. Among the four analytic rating scales, fluency explained 52% of the holistic rating, indicating its primary importance in assessing oral performance.

[3]Items under each factor were listed in the order of factor loadings, from the highest factor loading to the lowest one. For example, F2 in Data 1 had Items 6, 17, 11, 16 and 10. The loading value of Item 6 was higher than that of Item 17. Therefore Item 6 was put before Item 17.

Table 3.Interfactor correlation matrix

Factors	F1	F2	F3	F4
F1Fluency	1.000	-.599	.759	.446
F2 Conceptual understanding	-.599	1.000	-.532	-.266
F3 Communication clarity	.759	-.532	1.000	.460
F4 Communication appropriateness	.446	-.266	.460	1.000

Note. The notations are the same as those in Table 2.

Table 4.Total variance explained

Factors	Total	% of Variance	Cumulative %
F1Fluency	10.452	52.262	52.262
F2 Conceptual understanding	1.096	5.480	57.743
F3 Communication clarity	1.054	5.268	63.011
F4 Communication appropriateness	.499	2.495	65.506

Note. The notations are the same as those in Table 2.

4.5. What is the degree of model fit?

When conducting CFA to test model fit using the statistical software AMOS, I started with the model based on the pattern matrix that was developed through EFA, designated as Model 1 and shown in Table 5. This model was a 20-item factorial model. In the course of performing CFA, the items that had low loadings with other items within the same factor were deleted to improve the degree of model fit. For example, I deleted Item 7 in Model 2 and deleted Items 7 and 13 in Models 3 and 4. After several regroupings of these items and inspections of item loadings, I retained four models and compared their model fits. Table 5 shows a summary of goodness-of-fit statistics of the four models.

Model 1 was a four-factor solution with 20 items: F1 (Items 19, 9, 14, 3, 5, 4, and 1), F2 (Items 17, 6, 11, 16, 10), F3 (Items 18, 20, 12, 7), and F4 (Items 2, 13, 8). Model 2 was four-factor solution with 19 items, in which I deleted Item 7 due to its low loading. In Model 3 (four-factor solution with 18 items), I deleted Item 13 in addition to the deleted Item 7. In Model 4 (four-factor solution with 18 items), I deleted Items 7 and 13 and removed Item 1 from F1 to F3. The model fit indices were similar across the four models and Model 3 had the best model fit.

Table 5.Goodness-of-fit indices for the rating scale models

Models	$\chi^2_{S\text{-}B}/df$	*P*	GFI	CFI	RMSEA
Model 1	4.343	.000	.803	.890	.103
Model 2	3.890	.000	.827	.910	.095
Model 3	3.724	.000	.841	.922	.093
Model 4	4.382	.000	.826	.910	.103

In Model 3, the value of $\chi^2_{S\text{-}B}/df$= 3.724< 5.0, was lower than the upper threshold of 5.0, showing a good fit of the model. The corresponding *p* value was significant, which meant the fit was poor. However, this study had a large sample size (104*3 = 312). If a sample size was large enough, even small residual covariance associated with a well-fitting model may yield a significant *p* value, leading to a rejection of the model. Therefore, the *p* value in this study was not a good indicator for model fit. The CFI value of .922 (> .9) denoted an acceptable model fit, even though the value of .922 was close to the threshold of .9.The GFI value of .841 (< .9) did not signal an adequate model fit, but the value of .841 was close to the upper threshold of .9. Finally, the RMSEA value of .093was acceptable. It was < .1 and > .08, an indication of adequate fit according to Browne and Cudeck (1993).Overall, the values of $\chi^2_{S\text{-}B}/df$, GFI, CFI, and RMSEA did not demonstrate a strong model fit to the data, but the fit was acceptable.

5. Discussion

Differs from prior studies (e.g., Iwashita et al., 2008; Jin & Mak, 2013; Plough et al., 2010; Sawaki, 2007) that focus on linguistic components, the analytic rating scales developed in my study included both linguistic components and other components used for assessing communicative ability. The scale of "conceptual understanding" related to the linguistic component because two items clustered under this scale indicated "language accuracy." Three other items under this scale related to "conceptual understanding," which provided the locus for understanding linguistic utterances in contexts and incorporated pragmatic considerations. Shown in the results section, the scale "fluency" included the aspect of "content" and the ease of delivery. Both my study and Sato's (2012) confirmed the importance of the content component in assessing oral performance. In addition, the ease of delivery and other scales associated with "communication clarity" and "communication appropriateness" all denoted communicative ability.

Chinese teachers perceived "fluency" to be the most salient among the four scales. This result echoes findings of prior studies that recognize "fluency" as one of the deciding factors in determining raters' holistic judgments (Adams, 1980; Iwashita et al., 2008; Jin & Mak, 2013; Wang, 2002).

The four analytic rating scales developed in my study did not fully account for teachers'

holistic judgments of overall oral proficiency. In Sato's (2012) study, five scales accounted for 67% of raters' holistic judgments, which is similar to the results in my study (65.5%). This finding is not counterintuitive considering that language use always occurs in a context. Therefore, analytic rating scales that mainly reflect decontextualized structure of language ability should partially, rather than wholly, interpret oral proficiency. Other indigenous rating criteria or non-criterion features should be added to increase rating reliability. Many scholars, such as Douglas (2001), argue for using indigenous analytic rating scales to assess language proficiency considering specific target language situations.

The correlations among the four analytic rating scales were relatively high. The finding of high correlations among analytic rating scales is consistent with the results in Sawaki (2007) and Xi (2007). Although specific analytic rating scales and language contexts in the two studies of Sawaki (2007) and Xi (2007) differ from those in my study, the high correlations among analytic rating scales are similar to my study.

6. Implications

The results of the present study enrich our understanding of analytic rating scales used to assess L2 Chinese oral performance. This study demonstrates a way of determining the number of analytic rating scales, a finding that prior studies lack, which is a necessary step in developing rating scales for assessing oral proficiency. The study explores four analytic rating scales of "fluency," "conceptual understanding," "communication clarity," and "communication appropriateness." The content of the four scales hinge on the shared features of the items clustered under those scales. Different from previous studies (e.g., Iwashita et al., 2008; Jin & Mak, 2013; Plough et al., 2010; Sawaki, 2007) that adopt, rather than develop, analytic rating scales, this study provides a way of developing analytic rating scales for assessing oral proficiency. In addition, this study advances our understanding of the relationships between analytic rating scales and holistic judgments. This study explores the relative weights of individual scales that theoretical models of analytic rating scales have not explicated, as mentioned in the literature review section. The findings of this study readily apply to the process of developing construct definitions and scales for oral proficiency tests. All above findings provide Chinese teachers with a reference to assess U.S. college students' L2 Chinese oral performance.

The four scales explored in this study explain 65.5% of teachers' holistic judgments of oral performance. This finding is a useful guide for assessing L2 Chinese oral performance. In L2 Chinese assessment practice, classroom teachers or assessment professionals estimate student achievement and/or proficiency by averaging the scores on analytic scales or by weighting all components differentially. This practice assumes that a holistic score which reflects global proficiency can be completely explained through analytic rating scales. The results in this study, however, demonstrate that analytic rating scales could explain only 65.5% of raters' holistic judgments, rather than 100%. Therefore, this study suggests that teachers and assessment professionals need to add other facets to evaluate oral performance, such as students' self-descriptions of their own language ability or longitudinal records of teachers' ratings of these students.

7. Limitations

Some limitations of the present study should be noted. First, the degree of model fit was acceptable, but not strong. Second, many participants in this study have limited years of teaching experience. The participants included ten experienced L2 Chinese teachers when I interviewed them regarding their perceptions of analytic rating scales and rating items, 42 native speakers of Chinese in the pilot study, 156 L2 Chinese teachers in Study 1, and 104 L2 Chinese teachers in Study 2. The total participant number is 302. If the 302 participants had all been experienced L2 Chinese teachers, the result might have been more reliable. In spite of the limitations noted above, the results of the study are valuable. According to Chen (2014), experienced teachers do not necessarily rate oral performance differently from non-experienced teachers. Chen (2014) provided empirical evidence that teachers and non-teachers of Chinese assess student oral performance using a similar analytic-rating-scale model, which means that experience as a teacher does not necessarily outweigh culturally-influenced perceptions. Rather, experienced teachers might use a rating model similar to non-experienced teachers. Since identifying an adequate pool of experienced Chinese teachers is difficult, constructing a rating model based on inexperienced teachers is valuable in predicting oral proficiency.

References

ACTFL. (1987).*ACTFL Chinese Proficiency Guidelines*.

ACTFL. (1999). *The ACTFL Proficiency Guidelines*—Speaking Revised.

Adams, M. L. (1980).Five co-occurring factors in speaking proficiency. In J. Firth (Eds.), *Measuring spoken proficiency (*pp. 1–6). Washington, DC: Georgetown University Press.

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

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*(2), 239-257.

Bachman, L. F. & Palmer, A. S. (1981).The construct validity of the FSI oral interview. *Language Learning, 31,* 67-86.

----. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford Applied Linguistics.

Bachman, L. F. & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *Modern Language Journal, 70*(4), 380-390.

Browne, M. W. & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). London: Sage Ltd.

Canale, M. (1983). From communicative competence to communicative to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 2-27). London: Longman.

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

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.

Chen, G. (2014) Teachers' and non-teachers' perceptions of a Chinese learner's oral performances. *Journals of National Council of Less Commonly Taught Language, 16,* 57-85.

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

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

De Jong, J. H. A. L. & Van Ginkel, L. W. (1992). Dimensions in oral foreign language proficiency. In J. H. A. L. De Jong (Eds.), *The construct of language proficiency* (pp. 112–140). Philadelphia: John Benjamin.

Douglas, D. (2001). Language for specific purposes assessment criteria: where do they come from? *Language Testing, 18*(2), 171-185.

Douglas, D. & Selinker, L. (1992). Analyzing oral proficiency test performance in general and specific purpose contexts. *System, 20,*317–328.

----. (1993). Performance on a general versus a field-specific test of speaking proficiency by international teaching assistants. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 235–256). Alexandria, VA: TESOL Publications.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(2), 208–238.

Fulcher, G. (2003). *Testing second language speaking*. London: Pearson Education.

Fulcher, G. (2010). *Practical language testing*. London: Hodder Education.

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

Hanban/Confucius Institute Headquarters. (2009). *Test syllabus for HSK–Advanced level*. Beijing: The Commercial Press.

Higgs, T. & Clifford, R. (1982). The push towards communication. In T. V. Higgs (Eds.), *Curriculum, competence, and the foreign language teacher* (pp. 57–79). Lincolnwood, IL: National Textbook Company.

Hudson, T. (2005). Trends in assessment scales and criterion-referenced language assessment. *Annual Review of Applied Linguistics, 25*, 205-227.

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269-93). Harmondsworth, Middleesex: Penguin.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24–49.

Jackendoff, R. (2002). *Foundations of Language*. Oxford, New York: Oxford University Press.

Jin, T. & Mak, B. (2013). Distinguishing features in scoring L2 Chinese speaking performance: How do they work? *Language Testing*, *30*(1), 23–47.

Kondo-Brown, K. **(**2002).A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*, 3–31.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 3*, 387–417.

Luoma, S. (2004). *Assessing speaking.* New York: Cambridge University Press.

Ministry of Education & State Language Commission, the People's Republic of China. (2011). *Spoken Chinese proficiency grading standards and testing guidelines*. Beijing: Beijing Language and Culture University Press.

Plough, I.C., Briggs, S.L., & Bonn, S.V. (2010). A multi-method analysis of evaluation criteria used to assess the speaking proficiency of graduate student instructors. *Language Testing, 27* (2), 235-260.

Sato, T. (2012). The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing, 29*(2), 223-241. Retrieved from http://dx.doi.org/10.1177/0265532211421162

Saussure, F. D. (1959). *Course in general linguistics*. New York: Philosophical Library.

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

The Office of Chinese Language Council International. (2007). *Chinese language proficiency scales for speakers of other languages*. Beijing: Foreign Language Teaching and Research Press.

The Office of Chinese Language Council International. (2008). *International curriculum for Chinese language education*. Beijing: Foreign Language Teaching and Research Press.

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

Wang, J. (2002). A study of the scoring of three types of oral test items (in Chinese).*Chinese Teaching in the World, 4,* 63–77.

Weigle, S.-C. **(**1998). Using FACETS to model rater training effects. *Language Testing 15,* 263–287.

Xi, X. (2007). Evaluating analytic scoring for the TOEFL® Academic Speaking Test (TAST) for operational use. *Language Testing, 24*(2), 251-286.

Young, R. F. (1999). Sociolinguistic approach to SLA. *Annual Review of Applied Linguistics*, *19,* 105-131.

 ISSN 2476-5880

Appendix

Below is a questionnaire about your perceptions of English-speaking learners' oral performance. Please answer Question 1 below about your teaching background before taking this questionnaire. Thank you very much.

Question 1: Have you taught Chinese as a foreign language in the U.S.?

A: YES. I have taught it for_____year (s).
B: NO.

Questionnaire: After you watch the video(s), please show your perceptions of the speaker's speaking ability by placing an "X" for the appropriate category.

1. 他/她听力很好。He/she has a good listening comprehension.

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

2. 他/她性格很好。He/she has a good personality.

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

3. 他/她说得很流利。He/she speaks fluently.

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

4. 他/她的交际水平比较高。His/her communication proficiency level is high.

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

5. 他/她表达得很清楚。He/she expresses him/herself clearly.

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

6. 他/她的想法和中国人差别很大。His/her conceptual understanding is different from that of native speakers of Chinese.

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

7. 他/她能听得懂问题。He/she can understand the questions.

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

8. 他/她交流时很懂应对技巧。He/she employs an effective communication strategy.

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

9. 这些问题对他/她来说很容易回答。These questions seem easy for him/her to answer.

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

10. 他/她有很多语法错误。He/she makes many grammatical errors.

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

11. 他/她不太了解中国人的思维方式和语言行为。He/she does know Chinese conceptual structure and verbal behaviors.

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

12. 他/她不太理解被问的问题。He/she does not understand the questions being asked.

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

13. 他/她是一个很讨人喜欢的人。He/she is a person that others would like to deal with.

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

14. 他/她的回答,内容很丰富。The content of his/her answer is informative.

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

15. 他/她选词造句很恰当。He/she selects appropriate words and structures.

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

16. 他/她很多词用错了。He/she uses many words incorrectly.

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

17. 他/她对中国的认知很多是错误的。He/she perceives Chinese conceptual structure incorrectly.

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

18. 他/她听力不太好。His/her listening comprehension is poor.

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

ISSN 2476-5880

19. 他/她说的话信息量很大。He/she delivers much information.

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

20. 我不知道他/她在说什么。I do not understand what he/she says.

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree