Journal of Language and Translation
Volume 7, Number 1(13), (pp.25-41), 2017

Direct and Semi-Direct Validation: Test Takers' Perceptions, Evaluations and Anxiety towards Speaking Module of an English Proficiency Test

Houman Bijani [1] , Mona Khabiri [2*]
[1] Department of Foreign Languages and Literature, Science and Research Branch, Islamic Azad University, Tehran, Iran
[2] Department of English Language Teaching, Islamic Azad University; Central Tehran Branch, Tehran, Iran

Received: 16 January, 2017 Accepted: 21 March, 2017

Abstract
This research study employed a mixed-methods approach to investigate the test takers' perceptions and anxiety in relation to an English language proficiency test called Community English Program (CEP). This study also evaluated the direct and semi-direct modes for speaking module of this test. To this end, 300 English as Foreign Language (EFL) students were recruited in the study as test takers. They were invited to take the CEP speaking test using five tasks of Description, Narration, Summarization, Role-play and Exposition in both direct and semi-direct test modes. Their perceptions and evaluations of both test modes, through questionnaires, interviews and observations were examined. The results of the factor analysis revealed that test takers' evaluations of both direct and semi-direct speaking modes were quite similar, yet not exactly identical. On the other hand, although test takers' anxiety was shown influential, the findings showed that the most determining factor in test takers' oral performance was their capability level. Capability level was the main reason why some test takers out-performed the others. The findings also demonstrated that test difficulty identification was complex, difficult and at the same time multidimensional. The quantitative results displayed that the raters were scoring speaking performances differently; the qualitative results also provided logic for the reasons of these differences on the side of the test takers. Finally, the impact of test takers' gender differences on their perceptions was found nonsignificant.

Keywords: Direct oral assessment, English as foreign language (EFL) learners, English proficiency test, Mixed-method approach, Semi-direct oral assessment, Test takers' perceptions

INTRODUCTION

As pointed out by Bachman and Palmer (1996) one of the areas of difficulty in language testing has always been the measurement of speaking skill. Assessing speaking skill, according to Fulcher (2003), is challenging since there are many factors influencing our impression of how well someone can speak a language. One of the most important factors that greatly influences speaking assessment is the method or mode used to measure this skill (Winke, Gass, & Myford, 2012). An oral interaction may be evaluated by

*Corresponding Author's Email: Monakhabiri@yahoo.com

whether they are in progress during the interview procedure or they are based on tape-made oral interaction. Clark (1978) provides the basis for distinguishing two different types of speaking tests including direct and semi-direct oral language assessment modes. The direct Oral Proficiency Interview (OPI), according to Berstein, Van Moere, and Cheng (2010), is a face-to-face interaction between an interviewee and his/her examiner/interviewer. Clark (1975 as cited in Huei-Chun, 2007) introduced the term 'direct' to refer to the test procedures that replicate the situation in real-life communication. Such tests require test takers to use language in actual, real-life situations. Winke and Gass (2013) argue that face-to-face direct tests are more valid than other test modes including semi-direct tests in most circumstances because they are considered to be the unmarked form of interaction, whereas communicating by telephone and speaking into the microphone are the marked ones.

The term semi-direct oral test was first coined by Clark (1978) to describe the tests, which elicit active speech from test takers through tape recordings, printed text booklets, or any other non-human elicitation procedure, rather than through a face-to-face conversation with a present interlocutor. Tape-based oral tests may be the result of a direct test in which the test takers are interviewed in a face-to-face situation, or a semi-direct test in which the test takers do the test in a language laboratory giving responses to tape prompts (May, 2006). The Simulated Oral Proficiency Interview (SOPI) was created as an alternative to direct OPI for the sake of having more feasibility of administration while ensuring high reliability and validity measures (Knyon & Tschirner, 2000).

The degree to which such tests are valid and reliable alternatives to direct oral tests was investigated by Stansfield (1991). Stansfield in a comparative study of OPI and SOPI tests argue that SOPI proved itself to be a valid and reliable substitute for the OPI. In comparison of scores on the two kinds of tests, he reports Pearson correlations between 0.89 and 0.95. A large majority of students (86%) who preferred the live test felt nervous in the taped test. Moreover, a majority of them (90%) found the taped test more difficult.

Stansfield (1991) justifies that the reason why OPI and SOPI are highly correlated is perhaps because both tests do not let test takers represent their interactive skills fully. He argues that, even in the OPI test, both the examiner and the test taker believe that it is the test taker's responsibility to perform the talking. Thus, the kind of spoken language is not the mirror manifestation of natural talk in a real speech context. However, in contrast, Shohamy (1994) argues that high correlations between scores on the two different tests provide necessary but inadequate evidence for the substitution of either for the other. In other words, according to her, these two tests may not be measuring identical things. She further argues that it is required to measure the validity of them from various perspectives, not just through a simple correlation of their outcome scores.

Among a number of possible sources of rater disagreement that have been studied in the literature of speaking assessment (e.g., Fulcher, 2003; Gan, 2010; Van Moere, 2012), test takers perceptions and viewpoints are critical, which can also affect the validity of speaking tests. Language testers are often interested in the reactions of students to testing situations; thus, their perceptions are typically regarded as a central place in many Second Language Acquisition (SLA) research studies (Nakatsuhara, 2011). Test-takers have traditionally been associated with test face validity, thus were not commonly regarded as determining in test validation process (Bachman & Palmer, 1996). Consequently, test validation is left to experts who have received relevant training on test development and test analysis. However, the way test takers feel about a test determines whether the test is acceptable or not (Elder, Iwashita, & McNamara, 2002). This is due to the fact that if test takers have a negative attitude towards the test, then it is less probable that they perform their best which will definitely affect test validity accordingly. Consequently, Messick (1989 as cited in Elder, Iwashita, & McNamara, 2002) suggests incorporating test

takers' perceptions as an essential element of construct validity evidence.

Brown (1993) find a significant correlation between test takers' scores and their attitudes to the test factors in a way that less proficient test takers responded less positively than the more proficient ones. Scott (1986) and Zeidner and Bensoussan (1988) also in their studies find out significant positive correlation between test takers' attitudes and feelings with their performances in speaking tasks of various levels of difficulty. That is, those test takers who had more positive attitudes to particular tasks found them easier to take and thus they scored higher accordingly. Young and Milanovic (1992) recommend that test takers of low ability suffer more from anxiety than high ability test takers; however, they found a negative correlation between assessments of anxiety and Speaking test scores.

A key issue which has frequently been shown to influence learners' speaking performance assessment to a significant degree is the gender factor and gender-based perceptions and evaluations when scoring test takers' performance (O'Loughlin, 2002). There has been some research studies (e.g., O'Loughlin, 2002) on the effectiveness of gender on speaking assessment, which points out that male and female speech styles are somehow different. O'Loughlin believes that females are more collaborative and cooperative, whereas males are more controlling and uncooperative. Such claims could have substantial implications for speaking assessment since they show that communicative competence is gender dependent.

PROBLEM STATEMENT AND LITERATURE REVIEW

Considering literature around the field of second language education only few studies used a mixed-methods approach to investigate test takers' perceptions and evaluations of direct and semi-direct oral assessment tests. Also, no research could be found concerning the change of various elicitation techniques in oral testing prompts that may affect test takers' output and hence their scores. Thus, although the results of some studies (e.g., Stansfield & Kenyon, 1992) suggest different test performances on direct and semi-direct speaking interviews by test takers, it is not conclusive whether such differences are due to test mode or something else. On the other hand, the differences between male and female test takers' with respect to their perceptions and evaluations towards the two speaking test modes and the impact of which on their speaking performances was not investigated according to the literature.

Consequently, this study investigated test takers' perceptions and evaluations towards direct and semi-direct speaking tests through including both a qualitative and quantitative analytical approach to a clearer picture of their attitudes. Therefore, the following research questions were formed:

1-What are the test takers' perceptions and evaluations of each speaking test mode? Is there any significant difference between their perceptions and evaluations of the two test modes? (Qualitative and Quantitative data)

2-Is there any significant difference between male and female test takers regarding their perceptions and evaluations of the direct and semi-direct speaking tests? (Quantitative data)

METHODS

Participants

300 Iranian adult English as Foreign Language (EFL) students, including 150 males and 150 females (between 17 to 44 years old) participated in the study as test takers. The students were selected based on stratified random sampling from the ones studying at Intermediate, Upper-intermediate, and Advanced levels at the Iran Language Institute (ILI).

Instruments

The speaking test

The present study used the Community English Program (CEP) test to evaluate test takers' speaking ability under various language use situations using five tasks of Description, Narration, Summarization, Role-play and Exposition in both direct and semi-direct test modes. The reason for the selection of this test was that it was an internationally valid test for evaluating students' speaking of various levels. It could also evaluate various aspects of test takers' speaking in different contexts. The purpose of the speaking test was to measure the extent to which second language speakers could produce meaningful, coherent, and contextually appropriate responses to the following tasks.

Task 1 (Description Task) was an independent-skill task, which reflected test takers' personal experience or background knowledge to respond when no input was provided (Bachman & Palmer, 1996). Tasks 3 (Summarizing Task) and 4 (Role-play Task) reflected test takers' use of their listening skills to respond orally. In other words, the content for the response is provided for the test takers through listening. For tasks 2 (Narration Task) and 5 (Exposition Task) the test takers were required to respond to pictorial prompts including sequences of pictures, graphs, figures and tables.

The tasks of the CEP speaking test were all implemented via two modes of task delivery: (1) direct and (2) semi-direct. The direct mode was designed to use in an individual face-to-face approach (i.e., a single test taker speaking to an interlocutor-here a rater), whereas the semi-direct mode was designed to use in a language laboratory setting. Since one purpose of the study was to compare and contrast test takers' perceptions in relation to the tasks used in the study, the tasks of each test mode were analyzed based on features of task difficulty (Robinson, 2001), which are described here in details. For the purpose of comparability, both modes of the test consist of one-way exchanges (monologic) in which the test taker is required to communicate information in response to prompts from the rater. However, the role play allowed for a more authentic information gap activity in which meaning is negotiated between a test taker and an interviewer (dialogic). The tasks were also classified as either planned (allowing preparation time) or unplanned (eliciting spontaneous language). Planning time, according to Robinson (2001), affects language output to a high extent regarding both accuracy and complexity. Furthermore, tasks were distinguished as either open (allowing a range of possible solutions) or closed (allowing a restricted set of possible responses). Task classification was also done as being convergent (involving problem-solving for arriving at a particular goal) and those which are divergent (without specific goals, involving decision making, opinion and agreement). In this study, the only convergent task was the role-play. In another classification, tasks were classified regarding perspective dimension. This was to ask the test takers to do the tasks from their own first person perspective or another person's point of view third person perspective. Finally, tasks were classified regarding their immediacy dimension. This was to ask the test takers to speak using Here-and-now and There-and-then language structures.

The task types used in this study could be classified into two categories with respect to their difficulty levels based on the given factors above (Robinson, 2001). The following Table 1 gives the classification of tasks and their predicted difficulty levels.

Table 1

Table of Predicted Task Difficulty Classification

Dimension	Difficult (predicted)	Easy (predicted)
Openness	Close (limited response)	Open (free response)
Information exchange direction	Dialogic	Monologic
Language convergence / divergence	Convergent	Divergent
Language planning	Without planning time	With planning time
Perspective	3rd person point of view	1st person point of view
Immediacy	There-and-then	Here-and-now

Test takers' questionnaire

A questionnaire was used to elicit the test takers' feedback on both modes of the speaking tests through focusing on their perceptions, anxiety and evaluation of the speaking assessment quality. The questionnaire had originally been developed by Luoma (2004) consisting of five items; however, to make it more suitable for this study, it was modified thus the new version consisted of 17 and 13 items, for the direct and semi-direct test modes respectively, on a Likert scale to ascertain whether test takers' reactions differed significantly according to their characteristics or not. The revised questionnaire was used in English and the reliability and validity measures were obtained through statistical data analyses after running the questionnaire in a pilot study. The details of the pilot study for validating the questionnaire was not included in this article for the sake of keeping the brevity of the work.

Test takers' interview

The test takers participating in both modes of the speaking tests were invited for an interview session to help further clarify the data collected by the questionnaire.

PROCEDURE OF THE DATA COLLECTION

Prior to the administration of the speaking test, all the test takers were given the instruction guide so that they would be able to find out what they were expected to do in details. It is noteworthy to indicate that, along with providing the test takers with written instructions for the semi-direct mode of speaking assessment, the spoke version of the same instructions was provided on the tapes as well.

The 300 test takers participating in this study were divided randomly into two groups in a way that half of the test takers took first the direct and the other half the semi-direct test mode, and then they changed roles for the second half. The reason for not having all the participants performed in both modes of the speaking test was due to the fact that performance in one mode would most certainly affect their performance on the other mode through enabling them to get used to the typology of the questions and this would invalidate the findings of the study. At this stage the performances on both modes were audio-taped so that they could be rated retrospectively.

As a requirement of the study, which involved close observation of test takers' performances especially under both direct and semi-direct mediated modes, after the completion of all the test tasks, the test takers were given the questionnaire and were all asked for an interview session concerning their attitudes towards both test modes.

Data Analysis Method

Exploratory Factor Analysis (EFA) was used, for the quantitative data analysis, to analyze and identify the influential factors concerning the test takers' perceptions and evaluations of both test modes. Confirmatory Factor Analysis (CFA) was used to neutralize the influential effect of other loading items loaded in each factor. Besides, the use of CFA will only account for the determining items through getting the maximum loading of only those items loaded in each factor at a desired eigenvalue. ANOVA was used to identify

any significant differences among the factors identified in the exploratory factor analysis. For qualitative data analysis, a triangulation study consisting of a questionnaire, interview and direct observation by the researcher was used to allow the test takers to express their views freely. The qualitative collected data from the questionnaires, interviews and observations were analyzed through coding based on how they provided response to each item of the questionnaire.

RESULTS

The first research question

What are the test takers' perceptions and evaluations of each speaking test mode? Is there any significant difference between their perceptions and evaluations of the two test modes?

To answer the first research question, test takers' responses to the questionnaires of both test modes along with their interviews were analyzed and coded based on the similarity and differences of the way they responded each item of the questionnaire. The result of test takers' performance feedback analysis of their perceptions, feelings, effectiveness and evaluation, clarity and further development of the speaking assessment, obtained from the interviews and questionnaires, demonstrated that negative reactions to both test modes were mainly due to time limits for all speaking tasks. Many test takers found themselves unable to complete their answers and claimed that the test did not reflect their true ability. One of the test takers (highly proficient) commented that:

> If there were more time, I would be able to take notes of my ideas and be able to have a better performance. *(A female participant)*

This revealed that frustration and anxiety may have a negative influence on some test takers' performance. To raise the fairness of the test, on the side of the test takers and in spite of the adequacy of their responses, an increase of response time would seem fairer. A number of test takers expressed anxiety in both test modes. Some of the various sources of their anxiety are discussed briefly:

Regarding the *direct* speaking test mode, a majority of the test takers felt that the oral interview was rather the exact reflection of their speaking competence since they could observe a close correlation between their own abilities and the speaking assessment procedure. They further perceived direct speaking test as a low anxiety test providing a stress-free atmosphere.

> It was a relaxing atmosphere attending the speaking assessment training program. It helped me apply what I learnt during the course in my real-life [testing contexts] *(A male participant)*

With respect to the test takers' viewpoints, for the *semi-direct* speaking test, many indicated that they had little awareness and exposure to voice recording technique. For a majority of them, this was their first exposure to these facilities, thus it was expected that they would react almost negatively to such a so called intimidating situation. Some test takers expressed their dissatisfaction by the following comments:

> I prefer to talk to a person. [Because in that case you will feel more relaxed]. *(A female participant)*

> Speaking to people seems more natural and friendly. There is no [feedback speaking on a tape]. *(A female participant)*

> I [don't] like this testing format. I need to use nonverbal communication to better imply what I mean. *(A male participant)*

However, some individuals expressed anxiety with respect to the installment of the video re-

cording device and the voice recording tools for the semi-direct speaking test. Of course, this is something that cannot be eliminated from a research.

> Talking on a microphone while being watched on a camera made me a bit unsure about my responses and I was frequently diverted from the discussion path. *(A male participant)*

Regarding the tasks used in the study, some test takers stated that they were rather stressed-out due to their unfamiliarity with a number of tasks, e.g., Exposition task, which lowered their oral performance. Thus, they suggested that providing the test takers with enough information on the given tasks would enhance their performance ability.

> [That was the first time I was taking such tasks]. I never had to speak about a graph in my classes. I really felt nervous. *(A female participant)*

Similarly, the test takers rated task validity based on the facility of the task and whether they could perform well enough which once again reflects the importance of task familiarity and its effectiveness in their performance ability.

A few individuals expressed anxiety originating from them. They stated that they were anxious since they were uncertain of themselves and did not trust their abilities which caused anxiety accordingly. Some argued that if they had practiced harder, they would have performed better.

> [I'm scared that I can't act as well as other can. I don't think I'm as fluent as the other are]. *(A female participant)*

However, although the test takers expressed their fear to this testing mode, it is surprising to note that they strongly supported it as a novel innovative testing movement. This finding is relatively in line with that of Brown (1993) who found that students reacted positively to the speaking tests which were difficult and they felt unprepared for them and considered them to be valid speaking test instruments.

With respect to test takers' task enjoyment, although the statistical data showed that the test takers preferred the ones with which they were more familiar and performed better, the analysis of their post-test interviews seems to show that there is no one-to-one relationship between task facility and task enjoyment. The relative lack of consensus of their talks in this respect suggests that topic enjoyment is a more determining factor of task enjoyment rather than difficulty level. This finding is parallel with that of Fulcher (2003) who found that students prefer topic over task facility when they are assessed.

To identify the different aspects of the speaking test that the test takers were able to distinguish, an Exploratory Factor Analysis (EFA) was used to reduce the data and identify the influential factors involved in obtaining the test takers' perceptions and evaluations of the speaking assessment. Then, an ANOVA, based on the EFA, was run to observe the possible existence of any significant differences between the test takers' responses to the direct and the semi-direct speaking tests. Table 2 below displays the EFA to demonstrate the influential factors related to test takers' perceptions and evaluations of the direct mode of speaking assessment. The coding schemes that emerged during data elicitation were classified into various categories including positive and negative comments. It is worthy to indicate that the scree plot of the eigenvalues produced an elbow at the sixth eigenvalue. The first eigenvalue accounted for about 44% of the total variance. The direct oblimin was used as the estimation method of rotation in the principle axis factoring of factor extraction.

Table 2

Exploratory Factor Analysis of Test Takers' Perceptions and Evaluations of the Speaking Assessment Test (Direct Mode)

	Factor					
	1	**2**	**3**	**4**	**5**	**6**
1. I understood the testing instruction	-0.539	0.181	-0.446	**0.466**	-0.287	-0.35
2. I had enough time to think about to the questions before I spoke	0.161	-0.141	-0.326	0.374	-0.347	**0.63**
3. I had enough time to answer the questions	-0.108	-0.300	-0.536	0.078	0.086	**0.53**
4. The rater was easy to understand	0.198	0.113	0.253	**0.738**	-0.133	-0.07
5. The rater used understandable language	0.135	0.134	0.083	**0.557**	-0.161	-0.13
6. The rater's gestures helped understand the language	0.128	0.056	0.327	**0.733**	-0.022	0.08
7. It was difficult for me to understand what the teacher said because of his/her accent	0.246	0.132	0.383	**0.447**	-0.287	0.16
8. The speech at which the rater spoke was just right	-0.138	0.201	-0.088	**0.799**	-0.046	-0.06
9. If a different interviewer/teacher had done the interview, I would have done better	0.296	0.315	0.315	-0.386	-0.356	0.42
10. The rater was tense	0.386	**-0.533**	0.244	-0.172	0.342	-0.39
11. I was confused by the rater's language	0.128	-0.285	0.294	**0.664**	0.366	-0.43
12. The test was easy	-0.078	0.364	**0.762**	0.272	-0.016	-0.17
13. I was nervous before the direct test	0.356	**-0.769**	0.198	-0.138	0.377	0.16
14. I could talk to the rater easily	-0.202	**-0.618**	0.383	0.012	0.201	0.22
15. I think direct (interview) tests better evaluate your speaking proficiency	**0.970**	0.191	-0.151	0.049	0.029	-0.06
16. I think the topics selected for the tasks were suitable	**0.678**	0.285	-0.301	0.326	0.133	-0.16
17. I think I did well on the test	-0.379	0.552	0.055	0.038	**0.641**	0.24

Extraction Method: Principal Axis Factoring.

a. 6 Factor extracted.

It is evident from the Table 2 above that there were six determining factors loaded with an Eigenvalue greater than 0.4 showing that there were 6 significant factors in test takers' viewpoints of the direct speaking test. The questionnaire items loaded in each factor were marked in bold on the table as well. The loaded factors were named as the following:

Factor 1 "**Task Suitability**": shows to what extent and in which tasks the test takers enjoyed participation. Besides, it distinguishes whether or not the test tasks were suitable enough in eliciting

test takers' speaking proficiency. Item numbers 15 and 16 having the loadings of 0.67 and 0.97 were loaded in this factor.

Factor 2 "**Test Anxiety**": shows to what extent and in which tasks the test takers had more/less anxiety responding in the direct oral assessment test. Item numbers 10, 13 and 14 having the loadings of 0.53 and above were loaded in this factor. Although some individuals stated that they were rather anxious during the interview session, the negative loading on this factor shows that the majority of the test takers seemed to have low anxiety level with the speaking tasks and that they could talk to the raters in a stress-free atmosphere.

Factor 3 "**Test Facility**": shows to what extent and in which tasks the test takers faced more/less difficulty responding. Item numbers 12 having the loadings of 0.76 was loaded in this factor.

Factor 4 "**Test Instruction clarity**": shows to what extent the test takers understood what was intended in each test task and whether each task instruction was fully explained beforehand and whether or not they had any difficulty understanding what the interviewer intended to communicate and whether his/her accent was confusing. Besides, whether or not test takers could benefit from the interviewer's gestures to have a better understanding of language. Item numbers 1, 4, 5, 6, 7, 8, and 11 having the loadings of 0.44 and above were loaded in this factor. Here, a majority of them indicated that the instructions were just enough and the interviewers' accents and speech rate were quite understandable. They stated that the raters were rather quite understandable and the gestures were relatively helpful in understanding the task intention.

Factor 5 "**Self Evaluation**": shows to what extent the test takers felt satisfied with their own performance on the test. Item numbers 17 having the loadings of 0.64 was loaded in this factor.

Factor 6 "**Test Timing Sufficiency**": displays to what extent the time dedicated to each test task, both for planning and responding, was enough. Item numbers 2 and 3 having the loadings of 0.53 and 0.63 were loaded in this factor. Here, few test takers believed that the amount of time given was enough and they expressed that if they had more time, they would perform better. In fact, comment on lack of sufficient time was reflected more than any other comments by the test takers.

It is noteworthy to indicate that item number 9, asking the test takers whether they would get a different score if they were interviewed/rated by a different rater, was not loaded in any factor. This shows that a majority of the test takers did not feel like receiving a different score if they were interviewed/rated by a different rater. This indicates that the interview and rating was highly valid on the test takers' point of view. This finding is rather against that of Scott (1986) who found that a majority of the test takers expressed their agreement on receiving different scores if they were interviewed by another rater.

Afterwards, the scores of the six factors of the EFA were then used as dependent variables in an ANOVA test to identify whether there is a significant difference among the test takers' perceptions and evaluations about the direct speaking test or not. Table 3 displays the ANOVA analysis of test takers' perceptions to the direct speaking test.

Table 3
ANOVA Analysis of Factor Scores of Test Takers' Perceptions and Evaluations of the Speaking Test (Direct Mode)

	Sum of Squares	*df*	Mean Square	F	Sig.
Between Groups	346.811	5	69.421	3.695	0.008
Within Groups	1666.944	96	19.913		
Total	2013.721	101			

The finding shows that there was a significant difference with respect to test takers' responses to the questionnaire items as obtained by factor analysis.

EFA suffers from the point of view that the identification of the influential factors is regarded based on maximum loading amount of the items under each factor. However, this, by no means suggests that the other items do not have any effect on the loading of the item(s) loaded under a factor. In other words, in EFA, the loading effect of other items has not been neutralized but simply ignored. It must be argued that, however, all the other test/questionnaire items have their own loading effect on any particular item(s) loaded on a particular factor as well. Accordingly, in order to only evaluate the loading effect of the item(s) loaded on each factor and thus neutralize the loading effect of all the other items, a Confirmatory Factor Analysis (CFA) was run. CFA was performed using IBM Amos version 22.0. Table 4 below displays the CFA results demonstrating the influential factors related to test takers' attitudes and perceptions and evaluations about the ***direct mode*** of the speaking test. The model fit indices of CFI, TLI, RMSEA and SRMR display that the model used to obtain test takers' perceptions and evaluations of the speaking assessment test was a good and suitable model.

Table 4

Confirmatory Factor Analysis of Test Takers' Perceptions and Evaluations of the Speaking Assessment Test (Direct Mode)

	Factor					
	1	2	3	4	5	6
1. I understand the testing instruction				0.527		
2. I had enough time to think about to the questions before I spoke						0.718
3. I had enough time to answer the questions						0.627
4. The rater was easy to understand				0.838		
5. The rater used understandable language				0.651		
6. The rater's gestures helped understand the language				0.724		
7. It was difficult for me to understand what the teacher said because of his/her accent				0.573		
8. The speech at which the rater spoke was just right				0.827		
9. If a different interviewer/teacher had done the interview, I would have done better						
10. The rater was tense		-0.643				
11. I was confused by the rater's language				0.715		
12. The test was easy			0.821			
13. I was nervous before the direct test		-0.781				
14. I could talk to the rater easily		-0.736				
15. I think direct (interview) tests better evaluate your speaking proficiency	0.974					
16. I think the topics selected for the tasks were suitable	0.793					
17. I think I did well on the test					0.973	
CFI: 0.913 TLI: 0.896 RMSEA: 0.584 SRMR: 0.796						

Table 5 below displays the EFA administered to demonstrate the influential factors related to test takers' perceptions and evaluations of the ***semi-direct mode*** of speaking assessment. With

respect to the semi-direct test mode, the scree plot of the eigenvalues produced an elbow once again at the sixth eigenvalue. The first eigenvalue accounted for about 47% of the total variance.

Table 5

Exploratory Factor Analysis of Test Takers' Perceptions and Evaluations of the Speaking Assessment Test (Semi-Direct Mode)

	Factor					
	1	2	3	4	5	6
1. I could hear the tape well	-0.432	0.489	-0.240	**0.629**	-0.251	-0.039
2. I understood the testing instruction	0.069	-0.050	-0.127	**0.575**	0.216	-0.823
3. I had enough time to think about to the questions before I spoke	-0.143	-0.024	-0.670	0.024	**0.531**	0.151
4. I had enough time to answer the questions	0.410	-0.077	0.311	-0.333	**0.624**	-0.411
5. I could understand the language of the speaker well	0.215	0.164	0.261	**0.799**	0.026	0.042
6. I was confused by the language of the speaker	0.017	0.171	0.235	**0.708**	-0.175	-0.018
7. It was difficult for me to understand what the speaker in the tape said because of its accent	0.008	0.369	0.018	**0.632**	-0.322	0.349
8. The speed at which the speaker spoke was just right	0.004	0.437	0.081	**0.836**	-0.077	0.139
9. The test was easy	-0.116	0.154	**0.872**	0.037	-0.004	0.153
10. I was nervous before the semi-direct test	-0.003	**0.812**	0.084	-0.347	0.354	0.154
11. I think semi-direct tests better evaluate your speaking proficiency	**0.956**	-0.060	-0.094	0.140	-0.019	0.032
12. I think the topics selected for the tasks were suitable	**0.834**	0.129	-0.131	0.305	0.097	0.143
13. I think I did well on the test	0.023	-0.190	0.234	-0.260	0.389	**0.675**

Extraction Method: Principal Axis Factoring.

a. 6 factors extracted.

Similar to the direct test mode, six determining factors were loaded with an Eigenvalue greater than 0.4 which shows that there were 6 significant factors in test takers' viewpoints of the semi-direct speaking test. The questionnaire items loaded in each factor were marked in bold on the table as well. The loaded factors were named as the following:

Factor 1 "**Task Suitability**": shows to what extent and in which tasks the test takers enjoyed participation, besides, whether or not the test tasks were suitable enough in eliciting test takers' speaking proficiency. Item numbers 11 and 12 having the loadings of 0.83 and 0.95 were loaded in this factor.

Factor 2 "**Test Anxiety**": shows to what extent and in which tasks the test takers had more/less anxiety responding using the tape-mediated oral assessment format. Item number 10 having the loadings of 0.81 was loaded in this factor. Loading negatively on this factors show that the majority of the test takers seemed to

that the majority of the test takers seemed to have low anxiety level with the speaking tasks.

Factor 3 "**Test Facility**": shows to what extent and in which tasks the test takers faced more/less difficulty responding. Item number 9 having the loadings of 0.87 was loaded in this factor.

Factor 4 "**Test Instruction clarity**": shows to what extent the test takers understood what was intended in each test task and whether each task instruction was fully explained beforehand in the tape, and whether or not they had any difficulty understanding the tape-speaker's accent. Item numbers 1, 2, 5, 6, 7 and 8 having the loadings of 0.57 and above were loaded in this factor. Here, a majority of them noted that the instructions were just enough and the speaker's speech rater and accent was quite understandable. A majority of them also argued that they were not confused by the tape speaker.

Factor 5 "**Test Timing Sufficiency**": displays to what extent the time dedicated in the tape for each test task, both for planning and responding, was enough. Item numbers 3 and 4 having the loadings of 0.53 and 0.62 were loaded in this factor. Here, few test takers stated that the amount of time given was enough and they expressed that if they had more time, they would perform better.

Factor 6 "**Self Evaluation**": shows to what extent the test takers felt satisfied with their own performance on the test. Item number 13 having the loadings of 0.67 was loaded in this factor.

Afterwards, scores for the six factors of the EFA were then used as dependent variables in an ANOVA test to identify whether there is a significant difference among the test takers' perceptions and evaluations to the semi-direct speaking test. Table 6 below displays the ANOVA analysis of test takers' attitudes and perceptions to the semi-direct speaking assessment test.

Table 6

ANOVA Analysis of Factor Scores of test Takers' Perceptions and Evaluations of the Speaking Test (Semi-Direct Mode)

	Sum of Squares	***df***	**Mean Square**	**F**	**Sig.**
Between Groups	29.724	5	5.937	2.714	0.027
Within Groups	183.194	72	2.152		
Total	212.962	77			

Similar to the direct test mode, the finding of the table shows that there was a significant difference with respect to test takers' responses to the questionnaire items as obtained by EFA.

As already indicated for the direct test mode, to evaluate the loading effect of the item(s) loaded on each factor and thus to neutralize the loading effect of all other items, a CFA was run.

Table 7 displays the CFA results demonstrating the influential factors related to test takers' perceptions and evaluations of the semi-direct mode of the speaking test. The model fit indices of CFI, TLI, RMSEA and SRMR display that the model used to obtain test takers' perceptions and evaluations of the speaking assessment was a good and suitable model.

Table 7

Confirmatory Factor Analysis of Test Takers' Perceptions and Evaluations of the Speaking Assessment Test (Semi-Direct Mode)

	Factor					
	1	2	3	4	5	6
1. I could hear the tape well				0.641		
2. I understand the testing instruction				0.657		
3. I had enough time to think about to the questions before I spoke					0.635	
4. I had enough time to answer the questions					0.666	
5. I could understand the language of the speaker well				0.773		
6. I was confused by the language of the speaker				0.704		
7. It was difficult for me to understand what the speaker in the tape said because of its accent				0.672		
8. The speed at which the speaker spoke was just right				0.723		
9. The test was easy			0.857			
11. I was nervous before the semi-direct test		0.792				
13. I think semi-direct tests better evaluate your speaking proficiency	0.895					
16. I think the topics selected for the tasks were suitable	0.846					
18. I think I did well on the test						0.693
CFI: 0.934 TLI: 0.908 RMSEA: 0.561 SRMR: 0.783						

On behalf of the tasks, although the test takers were awarded higher scores in the Description task than the other ones, they felt that they had more anxiety dealing with it. The hypothetical reason for this conflicting outcome could be due to the fact that test takers, when answering questions of the Description task, feel like they are communicating with real people. Thus, most probably, that is why they found the task more stressful than the others. However, since they were more used to interview tasks, as the most typical speaking task, they had better performance than the others.

The second research question

Is there any significant difference between male and female test takers with regard to their perceptions and evaluations of the direct and semi-direct speaking tests?

With respect to gender differences and the variation of male and female test-takers' viewpoints on the speaking test, a Chi-square test was performed to ascertain whether there were any significant differences between male and female test takers on both modes of the speaking test – direct and semi-direct – based on their responses to the questionnaire and interview questions. The

outcome, as displayed in Table 8 below, demonstrated that there were no significant differences between the two genders. This finding is although consistent with that of O'Loughlin (2002) who found no significant differences between male and female test takers with respect to their perceptions of the direct and semi-direct speaking tests, is in contrast with the one found by Zeidner and Bensoussan (1988) who discovered that female test takers reacted more negatively to speaking tests.

Table 8
Male and Female Test Taker Viewpoints on the Direct and Semi-Direct Speaking Tests

Format		Value	*df*	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Direct	Pearson Chi-Square	0.153[a]	1	0.696		
	Continuity Correction[b]	0.001	1	0.980		
	Likelihood Ratio	0.153	1	0.695		
	Fisher's Exact Test				0.730	0.491
	Linear-by-Linear Association	0.148	1	0.700		
	N of Valid Cases	150				
Semi-direct	Pearson Chi-Square	0.536[c]	1	0.464		
	Continuity Correction[b]	0.134	1	0.714		
	Likelihood Ratio	0.537	1	0.464		
	Fisher's Exact Test				0.715	0.358
	Linear-by-Linear Association	0.518	1	0.472		
	N of Valid Cases	150				
Total	Pearson Chi-Square	0.601[d]	1	0.438		
	Continuity Correction[b]	0.267	1	0.605		
	Likelihood Ratio	0.602	1	0.438		
	Fisher's Exact Test				0.606	0.303
	Linear-by-Linear Association	0.591	1	0.442		
	N of Valid Cases	300				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 17.27.
b. Computed only for a 2x2 table
c. 0 cells (.0%) have expected count less than 5. The minimum expected count is 16.77.
d. 0 cells (.0%) have expected count less than 5. The minimum expected count is 36.51.

DISCUSSION AND CONCLUSIONS

The findings of this study demonstrated that test takers have preferences for specific sorts of test tasks in a way that some tasks are perceived to be easier/harder than the others. This finding is relatively in line with those of (Brown, 1993; Scott, 1986; Zeidner & Bensoussan, 1988) who found different perceptions with regard to task difficulty on the side of test takers. The results of the EFA revealed that test takers' evaluation of the direct and semi-direct speaking tests were quite similar, although not exactly identical with respect to the identification of item numbers loaded in each factor. Although test takers' individual differences among which anxiety in particular was shown to have been influential, at least on the test takers' point of view, the findings showed that the most determining factor on test takers' speaking has been their capability level which is various among test takers and that is the main reason why some test takers out-performed the others. This finding is fairly consistent with those of Young and Milanovic (1992) who found out that ability but not anxiety is a more important determining factor influencing test takers' Speaking scores. This finding provides evidence on Tarone (1983) Capability Continuum Theory which suggests capability is hetero-

geneous and that test takers vary in capability from one another.

The study also showed that through the combination of qualitative and quantitative approaches, a better realization of the research concept is achieved. The use of a mixed-methods approach could better provide the researcher with various angles of the study. This approach provided enough evidence concerning the ways test takers treated the various testing facets. As the quantitative results display that the two groups of test takers had different perceptions and evaluations of the two oral test modes, the qualitative results provided logic for the reasons of these differences on the side of the test takers. This could highly compensate for the shortcoming of the previous research that was solely dependent on either quantitative or qualitative methods. Although the application of both methods is time consuming, the provision of deep insight with respect to the validity and reliability of the assessment will be definitely worth it.

This study also represented that a valid test of speaking should consist of both direct and semi-direct tests in order to provide sufficient assessment evidence which is similar to what Nakatsuhara (2011) refer to in his study of test takers' speaking in which a combination of both approaches provide decision makers with the best conclusive decision on test takers' speaking levels.

The findings of the study also demonstrated that test difficulty identification is complex, difficult and at the same time multidimensional (Gan, 2010). However, test takers' perceptions could be considered as a reliable factor for determining task difficulty. However, this finding should not be misinterpreted as a key factor to establish a hierarchical order of task difficulty solely on the basis of test takers' testing intuitions. Generalizations also should be done with great caution.

Anxiety was identified as a significant influential factor in reducing test takers' performance. However, providing the test takers with adequate and explicit warm-up exercises prior to test administration, establishing a friendly atmosphere, and building confidence on test takers could definitely improve their performance to a considerable extent and compensate for the debilitative role of anxiety for test takers. This finding is fairly consistent with that of Van Moere (2012) who consider the role of anxiety as a determining factor in test takers' speaking.

References

Bachman, L. F., & Palmer, L. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Berstein, J., VanMoere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27*(355-377). doi: 10.1177/0265532207077205

Brown, A. (1993). The role of test-taker feedback in the test development process: Test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *1993, 10*(3), 26. doi: 10.1177/026553229301000305

Clark, J. L. D. (1978). *Direct testing of speaking proficiency: Theory and application*. Princeton: Educational Testing Service.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education*. London: Routledge.

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? . *Language Testing, 19*(4), 347-368.

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

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher- and lower-scoring students. *Language Testing, 27*(4), 585-602. doi: 10.1177/0265532210364049

Huei-Chun, T. (2007). *A study of task type for L2 speaking assessment*. Paper presented at the Annual Meeting of the International Society for Language Studies (ISLS), Honolulu, HI.

Knyon, D. M., & Tschirner, E. (2000). The rat-

ing of direct and semi-direct oral proficiency interviews: Comparing performance at lower proficiency levels. *The Modern Language Journal, 84*(1), 85-101. doi: 0026-7902/00/85-101

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

May, L. A. (2006). An examination of rater orientations on a paired candidate discussion task through stimulated verbal recall. *Melbourne Papers in Language Testing, 11*(1), 29-51.

Nakatsuhara, F. (2011). Effect of test-taker characteristics and the number of participants in group oral tests. *Language Testing, 28*(483-508). doi: 10.1177/0265532211398110

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing, 19*(2), 169-192. doi: 0.1191/0265532202lt226oa

Robinson, P. (2001). Task complexity, task difficulty and task production: Exploring interactions in a componential framework. *Applied Linguistics, 21*(1), 27-57.

Scott, M. L. (1986). Student affective reactions to oral language tests. *Language Testing, 3*(1), 99-118. doi: 10.1177/026553228600300105

Shohamy, E. (1994). The validity of direct versus semi- direct oral tests. *Language Testing, 11*(2), 99-123. doi: 10.1177/026553229401100202

Stansfield, C. W. (1991). A comparative analysis of simulated and direct oral proficiency interviews. In S. Anvian (Ed.), *Current developments in language testing* (pp. 199-209). Singapore: Regional English Language Center.

Stansfield, C. W., & Kenyon, D. M. (1992). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System, 20*(347-364). doi: 0436-251X/92

Tarone, E. (1983). On the variability of interlanguage systems. *Applied Linguistics, 4*(2), 142-164.

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing, 29*(3), 325-344. doi: 10.1177/0265532211424478

Winke, P., & Gass, S. (2013). The Influence of Second Language Experience and Accent Familiarity on Oral Proficiency Rating: A Qualitative Investigation. *TESOL Quarterly, 47*(4), 762-789. doi: 10.1002/tesq.73

Winke, P., Gass, S., & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing, 30*(2), 231-252. doi: 10.1177/0265532212456968

Young, R., & Milanovic, M. (1992). Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition, 14*(4), 403-424.

Zeidner, M., & Bensoussan, M. (1988). College students' attitudes towards written versus oral tests of English as a foreign language. *Language Testing, 5*(1), 100-114. doi: 10.1177/026553228800500107

Biodata

Houman Bijani is a PhD candidate in the field of Teaching English as a Foreign Language (TEFL) at the Islamic Azad University, Science and Research Branch, Tehran, Iran. He is also a faculty member of Zanjan Azad University. He was top student among other graduates in his master's degree in the field of TEFL (Teaching English as a Foreign Language) at the Allameh Tabatabai University. He has published several research papers in national and international language teaching journals. His areas of interest include quantitative assessment, teacher education and language research.
Email: houman.bijani@gmail.com

Mona Khabiri is an associate professor in the field of Applied Linguistics at the Islamic Azad University, Central Tehran Branch. She is also the director of the Journal of English Language Studies (JELS) in Iran. She mainly teaches the units in relation to language testing and assessment, research methodology, research seminar in the field of Teaching English as a Foreign Language (TEFL), and teaching language skills at graduate level. Her main areas of interest include teacher education, cooperative learning, and language testing and research methodologies. She has published papers in international and national academic journals and also presented in several national and international seminars.
Email: Monakhabiri@yahoo.com