An Analysis of the Differences between Spoken English Test Styles: Non-Response Tester VS Tape-Recorder

1. Introduction

If we classify the spoken English test (SET) according to the time a candidate is scored, we will probably have three types – direct, indirect and semi-direct SET. Direct SET is a traditional test type in which the assessor scores the performance of the candidates when they are in the test. In an indirect test, the performance of candidates is videoed or recorded by the tape-recorder and is scored later by the examiner. A semi-direct SET is a test type in which the examiner and candidates communicate face to face and the whole process is videoed or recorded for the examiner to do the assessment later rather than on the spot.

In his unpublished thesis for MA degree at Zhejiang University, Li (2000) listed the advantages and disadvantages of these three test types (Table 1). Based on this, Li concluded that:

- As for an achievement test, the authenticity and impact are even more important to some degree than reliability, direct SET is the favored format.
- Much effort, especially in the training of assessors, should be done to improve reliability of direct SET.
- However, the indirect SET deserves research effort so that a comparison can be made with the direct SET.

According to Guo Jieke (cited in Yang, 2002), assessors in the traditional SET, that is in the direct type, would make testees feel nervous. However, the testees do not feel so when they face a virtual examiner and then the real ability of the testees is expected. And now a computer-mediated SET system has been developed. In this system, the test is hosted by a virtual examiner who speaks American English and is scored by real assessors. The difficulty level of the computer-mediated test is the same as the traditional one, except that the former takes longer time. The computer-assisted type includes two more parts: picture description and picture comment. The purpose of these two more parts is to make the testees talk more on different types of questions besides the traditional ones, so that the result will be more objective. Finally, Guo says, because there is no problem over the number of examiners, the computer-mediated SET can afford to test much more students than the traditional direct SET.

While face-to-machine test becomes more and more popular, IELTS sticks to face-to-face form. IELTS requires candidates to be flexible and sensitive in real communication. It focuses on testees’ real ability to survive in English speaking countries, so communication is the main task in the test.

Now the case is rather complex. Because both face-to-face and face-to-machine test claim that its pole is the right way to test the real ability of examinees while the two poles are on the opposite site. On one hand, it is real examiners with communication hosting the test; on the other hand, it is virtual examiners without communication hosting the test. Then it seems interesting to find out what testees will react when they face a real examiner without communication and when they face a machine (tape-recorder).

Certainly, the virtual examiner will not have real communication with testees. The further questions or responses of the virtual examiner will not be as natural as real testers. And also, in the computer-mediated type, candidates do better because the virtual examiner is not someone who will laugh at the performance of them, or is
someone who looks like a human being, or because the examiner simply does not exist? If it is the third one, it seems that the virtual examiner acts no more than a tape recorder.

In this paper, we will try to find out the function of non-response human tester and tape recorder. The hypothesis is that testees will perform better when they face human beings, as it is more natural for them to talk. Because people tend to communicate with live things rather than machine, such as a tape recorder.

2. Research
2.1 Subjects
44 third year English majors, from College of Further Development, Zhejiang University. They were divided into two groups according to their student number: Group A number odd, Group B number even.

The students would not be informed of anything of the experiment, however, as the experiment was taken as the final test for the course of Spoken English, they were supposed to take the test seriously.

2.2 Materials
Because of the purpose of the experiment, the difficulty level of two trials in the same test should be the same. In the two trials, two aspects of a same case were presented to the testees (See Appendix). The two questions asked testees to solve problems for two persons in the same situation respectively. Though the case was identical, Material 1 and Material 2 were stated in different ways, focusing on the 23 years old woman and the 9 years old boy, so that testees would not ignore reading attentively.

2.3 Environment
Three rooms were used in the experiment, namely Room A, B and C.
Room A was where the tester sat. Testees would report their solution to the tester.
Room B was where the tape recorder was used. Testees would record their answers to the machine.
Room A and Room B were bright and in the same size.
Room C was used as a waiting room.
Both Room B and Room C had one assistant.
All students were asked to turn off any electronic communication devices to avoid the leakage of test content, as it was very probably that the students who had taken the test would inform later testees the content of the test by calling them from outside of classroom or sending short messages. If this happened, the reliability of the experiment would be a big problem.
To add to the security of the test reliability, both Room B and Room C would be under control of the assistant there. The assistant in Room B would hand the question sheet to testees, remind them of the time if they were over timing, and also make sure they do not rewind the tape to listen to others’ answers. The assistant in Room C would ensure that no one was obtaining any information about the content of the test before he/she should have known it. The assessor in Room A would not show any positive or negative response to the answer of testees.
Table 1 Advantages and disadvantages of three spoken English test types.

<table>
<thead>
<tr>
<th>Test types</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
</table>
| Direct SET   | • The assessor can observe facial expressions and body languages of the candidates, which is also considered factors attributing to communicative competence.  
• The assessor can remind the candidates if they speak too little or in a voice too low to be heard clearly.  
• The face-to-face communication enables relatively high authenticity and face validity.  
|              | • Training of a large number of assessors is expensive and time-consuming.  
• Examiners/assessors’ linguistic competence and affective factors (e.g. mood) may affect the performance of candidates.  |
| Semi-direct SET | • Semi-direct SET has all the advantages of direct SET.  
• The recorded tape can be reassessed to improve reliability.  |
| Indirect SET   | • The examiner does not affect candidates’ performance.  
• All the candidates receive exactly the same linguistic input, thus this may be fairer to them.  
• The assessment is not affected by candidates’ appearance.  |
|              | • Semi-direct SET has all the disadvantages of direct SET.  
• It is even more expensive and time-consuming in that the administration and assessment are done separately and thus the cost is doubled.  |
|              | • The laboratory setting and man to machine communication are not authentic.  
• Assessment is impossible in case that the candidate speaks very little or that the recording is unclear.  
• The assessment costs as much as direct SET in that the assessors have to listen to all the recordings of candidates.  
• The feasibility of this test format is dependent on accessibility of enough large language laboratories.  |

2.4 Procedure

The instruction of the test would be written on the blackboard in Chinese and explained to the students item by item. The detailed procedure is in Table 2.

Table 2 Classification of groups and steps of the test.

<table>
<thead>
<tr>
<th></th>
<th>Step 1</th>
<th>Step 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Type</td>
<td>Material</td>
</tr>
<tr>
<td>Group A</td>
<td>Tester</td>
<td>Material 1</td>
</tr>
<tr>
<td>Group B</td>
<td>Recorder</td>
<td>Material 2</td>
</tr>
</tbody>
</table>
For example, the first student in Group A (S1) would go to Room A, get Material 1 from tester, read, prepare and answer it. Then S1 would move to Room B, get Material B from the assistant there and record the answer to the machine. Testees in Group A would take Material 1 first, then Material 2. Testees in Group B Material 2 first, then Material 1.

In each trial, testees had one minute to read the material and prepare for it. They were allowed about three minutes to give their report. If testees take too long time, they would be stopped.

### 2.5 Scoring

In each trial, the score would be given according to three categories in Table 3, and the score description is in Table 4.

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency</td>
<td>10</td>
</tr>
<tr>
<td>Accuracy</td>
<td>10</td>
</tr>
<tr>
<td>Content</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Score Description</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>10</td>
</tr>
<tr>
<td>Good</td>
<td>8-9</td>
</tr>
<tr>
<td>Average or acceptable</td>
<td>5-7</td>
</tr>
<tr>
<td>Poor or failure</td>
<td>0-4</td>
</tr>
</tbody>
</table>

### 2.6 Data and Analysis

Before we start to analyze the data, it should be ensured that the result had not been affected by the order of materials, i.e. there was no practice factor involved in the experiment.

F test \( F_{(0.05/2)}=2.395, 1.085<2.395\) \(F<F_{(0.05/2)} P>.05\) shows that the variance of A1 and B2 are of no significant difference at .05 level, which allows further t test of mean difference.

\(t\) test \(t= 0.188; t_{0.05/2}=2.021> 0.188, P<.05; t_{0.01/2}=2.704> 0.188, P<.01\) shows that there is no significant difference both at .05 and .01 level. And thus we can safely say that testees did not get inspiration or practice from the previous trial. The final data are shown in Table 5. \(t_{0.05/2}=2.08.\) t values with one asterisk are significant at .05 level.)

### 2.7 Result

\(t\) test shows subjects did not react much differently when they were facing either non-response tester or tape-recorder. However, the test types did seem to affect the fluency of examinees. In both Group A and B, students appeared to talk more fluently with a recorder. This was totally opposite to what had been expected: subjects should behave better when they were with a real tester than a machine.
Table 5  Mean score of the test. The asterisk means the figure is significant at .05 score.

<table>
<thead>
<tr>
<th>Group A</th>
<th>Tester A1</th>
<th>Recorder A2</th>
<th>t test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency</td>
<td>6.23</td>
<td>7.14</td>
<td>2.83*</td>
</tr>
<tr>
<td>Accuracy</td>
<td>6.82</td>
<td>7.18</td>
<td>1.40</td>
</tr>
<tr>
<td>Content</td>
<td>7.14</td>
<td>7.23</td>
<td>0.28</td>
</tr>
<tr>
<td>Total</td>
<td>20.18</td>
<td>21.55</td>
<td>2.06</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group B</th>
<th>Recorder B1</th>
<th>Tester B2</th>
<th>t test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency</td>
<td>7.14</td>
<td>6.64</td>
<td>2.22*</td>
</tr>
<tr>
<td>Accuracy</td>
<td>6.95</td>
<td>6.77</td>
<td>1.07</td>
</tr>
<tr>
<td>Content</td>
<td>7.00</td>
<td>7.23</td>
<td>1.23</td>
</tr>
<tr>
<td>Total</td>
<td>21.10</td>
<td>20.64</td>
<td>1.19</td>
</tr>
</tbody>
</table>

3. Discussion

If students’ not behaving better in the experiment was due to the non-response strategy of the human tester, then IELTS seems to be right to stick to real human tester who can afford real communication. Thus on the other hand, the coming CET-SET computer-mediated type seems to have been losing its promising land: as inferred from the description of the coming system, the virtual tester is of no real response. And according to the above experiment, there was no significant difference in general when students were facing a non-response tester and a tape-recorder. Then it appears to be a waste of time and money to design software whose function in most places is already covered by a tape-recorder.

The result indicated that students could speak more fluently when they were with a tape-recorder. Was it because a recorder can not draw testees’ attention away as able as a real tester, for a recorder does not have much to observe while before a real tester, testees might have a lot to look at? Though the assessor kept similar facial expression and gave not a single positive or negative response, testees could still keep on wondering whether the tester was laughing at the their ability to speak English. A tape recorder does not have this problem, as examinees do not know when and where their report would be assessed, sometimes even who. While this sounds reasonable, there were 9 cases that subjects giggling or throat clearing before the recorder, and these are reactions when they are nervously speaking English to a real person. Does that mean subjects psychologically treat the recorder as an assessor sitting there? Or does that mean they were afraid that the assessor would laugh at their answers when the assessor listen to the tape? This needs further investigation.

Though result of this research shows that there is no significant difference between students’ performance when they were facing people and machine, there were some cases indicating that some people were extremely sensitive to different situations. Further experiments could be designed to see whether these differences are due to personal affective factors or simply caused by accident.

4. Improvement

About 1/6 of the students did not know how to operate the tape recorder. This seemingly tiny problem had slowed the whole process and more seriously maybe affect the performance of students, for this may cause much tension. Thus it is necessary to train testees to use the machine before the test.
When waiting for their turn, students in Room C seemed to have nothing to do except for chatting excitedly in mother tongue or dozing off or idling time away. These would probably make students not in the right mood at the beginning of the test, though in this experiment there is no significant practicing factor. However, it is suggestible to design proper activities for testees to do before test which will help them to be in the right mood.

5. Appendix

5.1 Material 1

You are a 23 years old young woman. One day when you are waiting for the bus, you notice a wallet on the ground. When you want to have a closer look, the man beside you picks it up and puts it into his bag. There are three people at the bus stop: you, the man who picks up the wallet and another man. The two men seem to know each other. Then comes a 12 years old boy who is looking for something. He comes to ask if you have seen his wallet. His description of the wallet is exactly the same as that of the wallet on the ground. The boy says his family needs money. And he used to cheat his father in order to use money to buy toys. His father was very angry with this when he finally realized what his son had done. The father gave the boy money this morning to buy books required by the teacher. The boy says his father will not believe him if he says he has lost the money but think he has used the money somewhere else. You feel you should tell him about the fact, but you are also afraid of those two men, because it is in the evening now. Then how will you cope with the situation?

5.2 Material 2

You are a 9 years old naughty boy who used to cheat your family in order to get money to buy toys you want. Your father had found this and was very angry with that because your family really needs money. This morning your father gave you some money to buy the books required by your teacher. But unfortunately you lost the wallet in the evening. You went back to the bus station and asked a young woman there. She said she did not see the wallet, but hesitated for a while. There were three people at the bus station including that young woman. You know someone in them must have taken your wallet. You know it is difficult for your father to believe the story, for you have cheated him before, though this time the story is true. Then how will you cope with the situation?
6. References


杨桂青, 2002, 英语四、六级口语考试：人机对话将成现实——访全国大学英语四、六级考试委员会副主任委员、华南理工大学教授郭杰克， 《中国教育报》 2002 年 9 月 18 日第 8 版。

张厚粲, 1993, 《心理与教育统计学》。北京：北京师范大学出版社。

作者信息

张璐
单位 浙江大学外国语言文化与国际交流学院
地址 浙江杭州 浙江大学紫金港校区 外国语言文化与国际交流学院
邮编 310058
email Najade@21cn.com