A Rasch analysis of an international English language testing system listening sample test

S. Vahid Aryadoust and Christine Goh

3rd Redesigning Pedagogy International Conference, Singapore, 1 - 3 June 2009

This document may be used for private study or research purpose only. This document or any part of it may not be duplicated and/or distributed without permission of the copyright owner.

The Singapore Copyright Act applies to the use of this document.
A RASCH ANALYSIS OF AN INTERNATIONAL ENGLISH LANGUAGE TESTING SYSTEM LISTENING SAMPLE TEST

S. Vahid ARYADOUST

Christine GOH

National Institute of Education, Singapore

Paper presented at the 3rd Redesigning Pedagogy International Conference

June 2009, Singapore
Title:

A Rasch Analysis of an International English Language Testing System Listening Sample Test

Abstract

This study reports on an investigation of the construct validity of an International English Language Testing System (IELTS) Listening sample test. The test was administered to 148 multinational participants. The Rasch modeling of data was used to fulfill the research objectives. Four major conclusions were made: 1) the Rasch differential item functioning analysis revealed that limited production items behave differently across different test taker groups suggesting the presence of construct-irrelevancies, 2) multiple choice questions (MCQ) do not cause construct-irrelevancies unless testees need to make ‘close paraphrases’ to comprehend the item stem or the question demands more than one answer; this nominates short MCQ as a best item format in listening tests, 3) evidence was found for ‘lexical processing’ which is different from top-down/bottom-up processing, and 4) the Wright map provided evidence for construct under-representation of the test. Findings from this study provide different sorts of evidence supporting and disproving the claim of the construct validity of the test, although they should be further investigated in future studies with different samples. Implications of the findings for IELTS and item writers are also discussed.

Key words: IELTS Listening, Rasch analysis, validity

---

1 An earlier version of this work has been presented in a proposal in partial completion of the degree of doctorate of philosophy in applied linguistics.
1. Introduction

The International English Language Testing System (IELTS) officially replaced English Language Testing Service (ELTS) in 1989. From this time on, extensive research was carried out on the test structure and modifications have been continuously applied. Nevertheless, many IELTS studies have been focused on predictive validity and washback effects or the productive skills of speaking and writing. To date, there is a paucity of research on the listening component of the IELTS test although a single published study by Coleman and Heap (1998) has tried to address this area (Aryadoust, forthcoming). Clearly, more research is needed, particularly on the way the construct of second language listening competence is accurately assessed. In view of the importance of examining this, the current study intends to investigate the (construct) validity of the listening component of the test via the Rasch analysis.

2. Concept of validity

Construct validity is whether or not the test measures what it is supposed to measure. After the emergence of the construct validity concept and its widespread investigation from the 1970s onward, evidence and use ultimately became the cornerstone of a new validation approach. Messick (1988), in his classification of validity, views the “source of justification of the testing” and the “function or outcome of the testing” as the main facets which prompt “evidence or consequence” and “test interpretation or use”, respectively.

In a further development of construct validity concept, Cronbach (1988) distinguished between two programs of construct validity: strong and weak. The strong program contains more explicitly stated theories whereas the latter is mainly “exploratory
empiricism” and focuses on validity in terms of any correlations of the scores with other variables. Cronbach believes that no interpretation of test scores and uses is definitive, ever-lasting, or absolute and that the more we progress in the science of psychology, the more the need to reconsider the interpretations. It is clear that the weak program proposed by Cronbach is very unifying but loose, lacking explicit instructions how to gather evidence, and “it is not clear that the strong program necessarily includes all kinds of validation efforts” (Kane, 2001, p. 326). Thus, a new argument-based approach has been proposed to investigate the validity of measuring instruments.

Messick (1989) introduced two factors that researchers and test developers must bear in mind in establishing the construct validity of tests. Construct-irrelevant factors are those off-target factors that can affect the test results. For instance, abilities to read and write in a listening comprehension test are construct-irrelevant (Buck, 2001). If these affect testees differently, then the prima facie evidence for construct-irrelevant factors exists. In addition, construct under-representation is a danger that may be posed to the construct validity of a test. If the test does not operationalize the construct sufficiently, it has fallen short in terms of measurement requirement to test the construct. Therefore, it behooves the test developer to take cognizance of these two detriments when operationalizing the construct and piloting the instrument. As will be reviewed below, Bond (2003) and Wright and Stone (1999) provide suggestions for analyzing validity of the tests by proposing two frameworks. These frameworks are Rasch-centered and will be employed in the current study.

3. Practical issues in construct validity
Messick (1994) proposes different statistical measures to inspect the construct validity of a measuring tool, such as factorial studies and correlation analysis. Additionally, Wright and Stone (1999) suggest that all validity classes (content, criterion, and construct) can be investigated through the Rasch analysis. They propose a model to explore the validity of the test. Presented below are a brief account of this model and an expansion of the Rasch measurement by Bond (2003) to meet Messick’s (1989) model.

3.1. The Rasch analysis

The Rasch model was proposed by George Rasch to analyze item difficulty and person ability. This method was introduced to offset the sample-dependent item properties traditionally computed in Classical Testing Theory (CTT). Accordingly, item difficulty (and person ability) will be sample-independent in the Rasch model. In this method, items and persons are calibrated on a same scale whose units are termed logit. For example, four items are located on a horizontal measure according to their difficulty levels (Figure 1). Persons attempting these items are also located on the same scale. The more the number of items and persons, the better the representativeness of the scale.

Figure 1

Relationship of persons and items calibrated on a common scale

Persons (A, B, C, and D) stand different chances to answer the items (1, 2, 3, and 4). For instance, for person C, it is possible to answer items 1 and 2. But item 3 is not
very probable for her to answer and item 4 seems very unlikely. This fact can also be showed on a curve known as Item Characteristic Curve (ICC) which is used in the Rasch modeling. In the Rasch we exploit the term probable which is simply why the Rasch is a probabilistic model. It is possible to show mathematically if a person’s ability to do the test is equal to the item difficulty, her chance of success on the item is 50% (see Wright & Stone, 1999). To find item parameters, Rasch models provide sophisticated and precise results based on mathematical models for the data. The basic model from which we can derive other models is:

\[
\phi_{ni} = \frac{\exp(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)}
\]  

(1)

where \(\phi_{ni}\) is person \(n\)’s probability of scoring 1 on item \(i\), \(\beta_n\) is the ability of person \(n\) on the entire test, and \(\delta_i\) is the difficulty level of item \(i\). Obviously, from this model, the probability of success to answer the question correctly is governed by person ability and item difficulty.

### 3.2. The Rasch Validity

Bond (2003) presents an argument to associate the Rasch and the two validity detriments (construct irrelevant factors and construct under-representation) proposed by Messick (1989). In this section, this argument is briefly discussed and Wright and Stone’s (1999) Rasch validity is covered. Prior to entering any discussion, it is necessary to state that Cronbach (1988) believes content and criterion validity are attempts to establish construct validity. Messick (1975) also does not agree that content-referenced evidence is any proof for validity; rather construct validity is the essential property of any measure. He states “Call it ‘content relevance,’ if you will, or ‘content representativeness,’ but do not
call it ‘content relevance,’ because it does not provide evidence for the interpretation of
the responses or scores” (p. 961). Following this argument, the terms validity and
construct validity are used interchangeably in the current article. Therefore, the ‘Rasch
validity’ is a plan to establish the (construct) validity of the measuring tools. This plan
considers validity as a multi-faceted property which should be investigated via different
methods which are Rasch-centered. The efficacy of this plan will be investigated below.

In the Rasch analysis, it is imperative to investigate the Wright (Item-Person)
Map to examine Messick’s (1989) proposed threats to validity. Item difficulty and person
ability indexes are calibrated on this map. Insufficient number of items to measure
persons of all ability ranges is a sign of construct under-representation (Bond, 2003).
Additionally, the fit of the data into the Rasch model is explored. Inspecting item and
person fit indexes (infit and outfit) will provide information about another cell of
Messick’s proposed threats to construct validity, which is construct-irrelevant factors.
Outfits are “heavily influenced by outlying, off-target, unexpected responses.” On the
other hand, “A useful alternative is to weigh residuals by the information they contain so
that the fit statistics are information weighted or “infits” and hence focus on inlying, on-
target, unexpected responses” (Wright & Stone, 1999, p.53).

Misfit patterns may be due to erratic typography read as errors, which may cause
higher item difficulties, guessing, which comes to play when students encounter an item
with a difficulty beyond their ability level, and carelessness, which is observed when high
ability students forget or fail to answer a simple question. In the meantime, “when the
disturbance in a misfitting item is not mechanical or clerical, the cause is usually special
knowledge” (Wright & Stone, 1999, p. 117). That is to say, if there is an interaction
between special students and some items, these items are unfairly biased in favor of/against these students. If the fit is too high, this means that high-skilled students are treated unfairly and if the fit is too low, this is the indicator that the item functions unfairly toward low-ability students.

Figure 2
Illustration of validity concept in the Rasch model according to Wright & Stone (1999)

The aforementioned discussion was intended to link Messick’s (1988, 1989) expositions of validity and threats to it to the Rasch analysis. Finally and in an effort to further explain how the Rasch model is linked with the validation procedures, the validity model of Wright and Stone (1999) is discussed here. Figure 2 illustrates the assumptions Wright and Stone (1999) make about the concept of validity in the Rasch. Validity in this illustration subsumes fit and order validity. Order validity has two subclasses: meaning validity, originated from the calibration of test items and utility validity, which is based on the calibration of persons. Wright and Stone go further to state that “The difficulty order of items defines the variable’s meaning and hence its content and construct validity” (p. 168).
Furthermore, from 1) the analysis of residual, i.e., the difference between the model and data, 2) analysis of item fit, which can help revising the test, and 3) analysis of person fit, which can help diagnosing the testees whose performance does not fit our expectations, we get response, item function, and person performance validity, respectively. These three validity types constitute another class of validity entitled the fit validity by Wright and Stone. Thus, this is the function of fit validity and order validity that establishes the validity of the test or tool. It is worth noting that Cronbach (1988) states that criterion and content validity are attempts to establish construct validity and therefore what is termed validity of the test scores is in fact the concept of construct validity. The Rasch model for validation described above is an attempt to establish this validity via different methods.

4. Objectives of the study

The specific aims of this study are to examine the following areas:

1. To determine dimensionality and local independence of the test.
2. To investigate construct-underrepresentation and construct-irrelevancies in the test.
3. To explore the degree of the operationalized construct equivalence across age, gender groups, education levels, nationality groups, and students with/out the experience of taking IELTS preparation courses.
4. To investigate if fit/order validity framework provides sufficient information to decide on the validity of the test.
5. Method

5.1. Procedures

Initially, consent forms (Appendix 1) were given to a randomly selected multinational sample of students who agreed to voluntarily enter the experiment, followed by an IELTS Listening Test. The majority of students were from China, Iran, Malaysia, and the Philippines (in addition to a small portion from other nationality backgrounds). First, participants were exposed to a short introduction into the IELTS Listening Module. The main objectives of the lesson are outlined below.

1) To introduce the IELTS Listening Module.

2) To review the item types and the test sections.

3) To review the rules that participants must know, such as writing no more than the assigned number of words in limited-production items, etc.

4) To give them an abridged listening test sample to practice the rules they have been taught. This test sample was selected from an IELTS practice material.

The main section of the study, which included a 40-minute IELTS Listening Test, was initiated after a break which was meant to eliminate testees’ probable fatigue caused in the introduction lesson. A questionnaire was given to them following the test (Appendix 2). The purpose of this tool was to collect information about the participants’ demographics to use in data analysis. All materials were collected in at the end of the study. The following table summarizes data collection procedures.

Table 1

<table>
<thead>
<tr>
<th>Activity</th>
<th>Time</th>
<th>Remarks</th>
</tr>
</thead>
</table>


<table>
<thead>
<tr>
<th></th>
<th>Duration</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instruction and practice</td>
<td>20 min.</td>
<td>Students received instructions; then, an abridged practice test was given to them.</td>
</tr>
<tr>
<td>A short break</td>
<td>10 min.</td>
<td>The break lasted 10 min.</td>
</tr>
<tr>
<td>Listening test</td>
<td>40 min.</td>
<td>30 min. for the test; and 10 min. for transferring answers to answer sheets</td>
</tr>
<tr>
<td>Questionnaire</td>
<td>5 min.</td>
<td>Termination of the study</td>
</tr>
</tbody>
</table>

5.2. Participants

The participants in this study were randomly selected from the students in the British Council (Malaysia branch), Singapore’s National Institute of Education (NIE), and a language institute in the Philippines. Ultimately, 148 people agreed to voluntarily attend the study. After the study, a short report showing their performance and their score was e-mailed to each individual.

5.3. Stimuli

A retired test of IELTS Listening chosen from Official IELTS Practice Materials was given to the students. The test comprises four sections. Sections one and three include two conversations between two inter-locutors. Whereas section one is focused on the survival matters of life, section three stimulates an interaction on a university campus. By the same token, sections two and four are respectively about daily matters and academic affairs. These two parts are mini-talks which are delivered by an expert in a field or someone who is introducing a merchandise or service. Each section is followed by 10 questions, so there are 40 questions in the entire test. Prior to answering questions in each section, participants are given some time to read the questions ahead. This creates a

---

2 Readers are referred to Official IELTS Practice Materials (2007) published by Cambridge University Press. Unfortunately, due to copy right restrictions, it is not possible to append the test items.
“purpose” that helps the listener answer the questions (Buck, 1991). At the end of the test, the test takers are given 10 minutes to transfer their answers to an answer sheet provided for them.

Although each question carries the same weight in the final raw score, from one test administration to another the task of translating the number of correct answers into band scores may vary. That is to say, for each test of listening the score-to-band conversion criteria can differ according to the test. This conversion regulation is kept confidential to the public.

6. Findings
6.1. Fit, dimensionality, and local independence

The Rasch analysis of items, persons, and residuals was conducted to fulfill the objectives of the study. WINSTEPS version 3.57 was used to carry out this analysis. In order to estimate parameters in the Rasch model, WINSTEPS uses the maximum likelihood estimation method. Parameter estimation is carried out in statistics through a lot of procedures such as least squares, and maximum likelihood. In the Rasch analysis, maximum likelihood, developed by Ronald Fisher, is used. The estimation of items and person parameters simultaneously is termed UCON, unidimensional maximum likelihood estimation (MLE). This method is presented in an implicit equation. Implicit equations are solved by using a method invented by Isaac Newton. One must guess the closest value for the unknown value and solve the problem. The proximity of this value to the most suitable solution is determined by exploring “how much remains when this value for X is substituted in the equation” (Wright & Stone, 1999, p. 124). This step may be repeated to arrive at the best solution. These steps are called iteration.
In the present study, item and person reliability indexes from the Rasch analysis were 0.97 and 0.91, respectively. These statistics are considered significantly high according to Bond and Fox (2007) and prove the sufficient Rasch reliability of the measurement. Respective separation indexes of persons and items are 3.2 and 5.7. This index is “the ratio of “true” variance to error variance” (Linacre, 2008, p. 462). This is another expression of reliability which ranges from 0 to infinity and indicates the number of performance levels in the test.

Following this observation, item measures (item difficulty) and fit indexes were requested (Table 2; Figure 3). Item measures are different from the observed raw scores in that they are converted into interval scales. These measures are then displayed on a map known as the Wright or Item-Person map. The Wright map is displayed to show the relative relations between item and person estimates visually. On this map, items are further analyzed to inspect whether or not they

<table>
<thead>
<tr>
<th>Item</th>
<th>Measure</th>
<th>S.E.</th>
<th>Infit MNSQ</th>
<th>Outfit MNSQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1.96</td>
<td>.28</td>
<td>.79</td>
<td>.53</td>
</tr>
<tr>
<td>2</td>
<td>-2.04</td>
<td>.29</td>
<td>1.02</td>
<td>.86</td>
</tr>
<tr>
<td>3</td>
<td>-.67</td>
<td>.24</td>
<td>1.44</td>
<td>2.00</td>
</tr>
<tr>
<td>4</td>
<td>-.12</td>
<td>.23</td>
<td>1.04</td>
<td>.93</td>
</tr>
<tr>
<td>5</td>
<td>-.01</td>
<td>.23</td>
<td>1.27</td>
<td>1.40</td>
</tr>
<tr>
<td>6</td>
<td>.96</td>
<td>.24</td>
<td>1.29</td>
<td>1.28</td>
</tr>
<tr>
<td>7</td>
<td>-1.13</td>
<td>.25</td>
<td>.85</td>
<td>.89</td>
</tr>
<tr>
<td>8</td>
<td>-.85</td>
<td>.36</td>
<td>1.23</td>
<td>3.49</td>
</tr>
<tr>
<td>9</td>
<td>-.73</td>
<td>.34</td>
<td>1.26</td>
<td>3.95</td>
</tr>
<tr>
<td>10</td>
<td>-.55</td>
<td>.24</td>
<td>1.02</td>
<td>.90</td>
</tr>
<tr>
<td>11</td>
<td>.09</td>
<td>.23</td>
<td>1.20</td>
<td>1.12</td>
</tr>
<tr>
<td>12</td>
<td>-2.13</td>
<td>.29</td>
<td>1.11</td>
<td>.97</td>
</tr>
<tr>
<td>13</td>
<td>-2.73</td>
<td>.34</td>
<td>1.06</td>
<td>.78</td>
</tr>
<tr>
<td>14</td>
<td>-1.88</td>
<td>.28</td>
<td>1.08</td>
<td>3.04</td>
</tr>
<tr>
<td>15</td>
<td>-1.19</td>
<td>.25</td>
<td>1.05</td>
<td>1.31</td>
</tr>
<tr>
<td>16</td>
<td>-.50</td>
<td>.23</td>
<td>1.28</td>
<td>1.58</td>
</tr>
<tr>
<td>17</td>
<td>.36</td>
<td>.23</td>
<td>.91</td>
<td>.81</td>
</tr>
<tr>
<td>18</td>
<td>-.12</td>
<td>.23</td>
<td>.95</td>
<td>.89</td>
</tr>
<tr>
<td>19</td>
<td>-.50</td>
<td>.23</td>
<td>1.51</td>
<td>2.36</td>
</tr>
<tr>
<td>20</td>
<td>.20</td>
<td>.23</td>
<td>.99</td>
<td>.94</td>
</tr>
<tr>
<td>21</td>
<td>.25</td>
<td>.23</td>
<td>.88</td>
<td>.79</td>
</tr>
<tr>
<td>22</td>
<td>-.72</td>
<td>.24</td>
<td>.81</td>
<td>.75</td>
</tr>
<tr>
<td>23</td>
<td>.80</td>
<td>.24</td>
<td>.76</td>
<td>.61</td>
</tr>
<tr>
<td>24</td>
<td>2.52</td>
<td>.30</td>
<td>.94</td>
<td>.66</td>
</tr>
<tr>
<td>25</td>
<td>-.07</td>
<td>.23</td>
<td>1.09</td>
<td>1.08</td>
</tr>
<tr>
<td>26</td>
<td>1.63</td>
<td>.26</td>
<td>.93</td>
<td>.70</td>
</tr>
<tr>
<td>27</td>
<td>1.83</td>
<td>.26</td>
<td>.99</td>
<td>.95</td>
</tr>
<tr>
<td>28</td>
<td>-.61</td>
<td>.24</td>
<td>.80</td>
<td>.65</td>
</tr>
<tr>
<td>29</td>
<td>.85</td>
<td>.24</td>
<td>.71</td>
<td>.57</td>
</tr>
<tr>
<td>30</td>
<td>-.17</td>
<td>.23</td>
<td>.84</td>
<td>.69</td>
</tr>
<tr>
<td>31</td>
<td>.47</td>
<td>.23</td>
<td>1.12</td>
<td>1.09</td>
</tr>
<tr>
<td>32</td>
<td>1.14</td>
<td>.24</td>
<td>.82</td>
<td>.65</td>
</tr>
<tr>
<td>33</td>
<td>.52</td>
<td>.23</td>
<td>.70</td>
<td>.56</td>
</tr>
<tr>
<td>34</td>
<td>1.38</td>
<td>.25</td>
<td>.76</td>
<td>.65</td>
</tr>
<tr>
<td>35</td>
<td>.25</td>
<td>.23</td>
<td>.73</td>
<td>.65</td>
</tr>
<tr>
<td>36</td>
<td>2.61</td>
<td>.30</td>
<td>.95</td>
<td>.56</td>
</tr>
<tr>
<td>37</td>
<td>-.17</td>
<td>.23</td>
<td>.70</td>
<td>.58</td>
</tr>
<tr>
<td>38</td>
<td>3.01</td>
<td>.33</td>
<td>.99</td>
<td>.64</td>
</tr>
<tr>
<td>39</td>
<td>2.61</td>
<td>.30</td>
<td>.92</td>
<td>.54</td>
</tr>
<tr>
<td>40</td>
<td>1.38</td>
<td>.25</td>
<td>1.12</td>
<td>.93</td>
</tr>
</tbody>
</table>

Mean 0.00 0.26 1.00 1.11
S.D. 1.48 0.04 .2 .78

Table 2
Item measures and fit indexes of the IELTS Listening Test
cover the domain intended to be measured according to the candidates’ ability. Item measures together with other Rasch indexes are shown in Table 2.

Overall, the item difficulty indexes range from -2.85 (Q8) to +3.01 (Q38) which denotes that the easiest question is Q8 in section one and the most challenging one is Q38 in section four. But person measures have a wider distribution and are homogenous. They range from -2.68 to +3.81, which indicates the sufficient utility validity of the sample. Another important observation is that on the whole questions in sections one and two are easier than the items in sections three and four.

Figure 3
Wright map of items and persons calibrated measures

To investigate the question of construct under-representation, Bond (2003) innovatively proposes scrutinizing the Wright map. If there are large gaps between bulks of items, one has some evidence for construct under-representation. Figure 3 displays
person and item measures on a Wright map. According to Figure 3, person measures are spread more homogenously than items measures. In addition, 7 gaps have happened among item measures on the map. A closer scrutiny shows that these gaps can be divided into two categories, i.e., major gaps which include gaps 1, 3, 6, and 7, and minor gaps which include gaps 2, 4, and 5. Apparently, the listening test is not representative of the listening construct for some test-takers in the current study who have the ability ranges from -2.5 to +2. This finding shows that meaning validity of items is not established.

Another consideration is to analyze items according to their difficulty. As Coleman and Heap (1998) state, based on the IELTS Listening Test guidelines “the items become more difficult as candidates proceed throughout the test” (p. 68). So, we expect to find the most difficult items in sections 3 and 4. However, examining the Wright map reveals that items 24, 26, 27, and 18 from section 2 and item 6 from section 1 are located at the top rank. There is a common core to all of these questions. All of them ask students to choose more than one alternative or write more than one word. For instance, item 6 requires that testees select two options out of six available alternatives and item 18 asks students to write no more than three words. An analysis of the responses to item 18 revealed that this item has been answered in three words by the majority of candidates. This is perhaps due to the rubric which reads: “Write NO MORE THAN THREE WORDS AND/OR A NUMBER for each answer.” It is possible that the test takers had thought the rubric demanded they write not less than three words or that a three-word response is the best answer. About LP questions in IELTS Listening, Coleman and Heap (1998) state that “student responses indicated they were unsure of how many words to write” (p. 65). Interestingly, it was further observed that many
candidates had written the first two words correctly (*four course dinner*). This observation will be further discussed in due course below.

Likewise, items 24, 26, 27, and 29 require that candidates write a phrase (24: *pink slip*; 26: *information desk*; 27: *25 pence*). Many testees had had problems recognizing the phrase ‘pink slip’ in Q24. Such answers as ‘pen split, fin slip, and pink click’ were chiefly observed among wrong answers. This issue was also observed in wrong answers to Q29 (correct: *essay plan*; incorrect: essay plane/assay plain or other combinations).

However, Q25 whose answer is a one-word response (departmental) is not difficult although it is located among this group of difficult items on the test paper. The reason can be due to the repetition of its answer twice in the aural passage (see Buck, 1991). It was further observed that many candidates had picked up the digit part of the answer to question 27 correctly, but as if the word ‘pence’ had been unknown to them, other counterparts such as cent, dollar, and pound are written.

It was also found that although questions 32, 36, 38, 39, and 40 are very difficult items, item 37 does not behave in accord with this group in which it belongs. The grouped questions demand answers comprising more than one word (like other items mentioned above), whereas the answer to this question is a single word (*die*). This finding is similar to the conclusion about Q25, which has a one-word answer and is located among more difficult questions on the test paper. Item 36 has clearly been troublesome in terms of its answering demands. Whereas the correct answer is the phrase “*leave the nests*”, such wrong answers as ‘live the net, leave net, and level their net’ were found dominant. The last finding to report is that the 10 items located at the bottom of the map are all from sections 1 and 2.
In an effort to investigate the fit of the data with the Rasch model, fit indices were examined (see Table 2). Finding misfitting (underfit) items indicates that either the item has been poorly designed or that it can be good on its own and not on the test investigated. Namely, the item may “not form part of a set of items which together define part of a single measurement... two or more measurement traits may have been confounded in the construction of the test” (McNamara, 1996, p. 175). But overfitting items are redundant and do not provide any extra information rather than other items. “Acceptable values [of standardized fit] fall between -2.0 and +2.0 with sample sizes between about 30 and 300” (Bond & Fox, 2007, p. 43). Bond and Fox (2007) also consider 0.6—1.4 as an acceptable mean square infit range. Likewise, Linacre (2005, pp. 215-217) recommends the following levels for infit (a $t$-standardized information-weighted mean square statistic which is more sensitive to the unexpected behavior of items closer to persons’ measures) and outfit (a $t$-standardized unweighted mean square sensitive to outliers):

a) Mean Square infit (MNSQ) >2.0, noise is more than information (underfit);

b) Mean Square infit (MNSQ) >1.5—2, some noise which is neither helpful nor degrades the measurement;

c) Mean Square infit (MNSQ): 0.5—1.5, productive of measurement (overfit);

d) Mean Square infit (MNSQ): <0.5, less productive, yet not degrading the measurement.

e) Mean Square outfit (MNSQ): substantially less than or above 1 values are very problematic.
More precisely, Pollitt and Hutchinson (1987) argue that the acceptable mean square infit statistics fall within the scope of mean ± twice the SD. In the current study, this value for items is in agreement with Bond and Fox’s (2007) proposed range, i.e., 0.60—1.40, and for persons is 0.63—1.35. Item and person measures which fall out of these ranges indicate variability not predicted by the model, which is detrimental to fit validity. Three persons’ infit indexes (2.7%) were slightly above the Pollitt and Hutchinson’s recommended range, yet still within Linacre’s proposed range (1.37, 1.38, and 1.42). Owing to the fact that, these infit mean square indexes do not show significant misfits, we are safe to assume that, according to Wright and Stone (1999), person performance validity is established. Bond and Fox (2007) point out that “Infit and outfit mean square values are always positive (i.e., > 0)… In this form, the mean square fit statistics are used to monitor the comparability of the data with the model” (p. 239). Therefore, a fit value of 1.25 indicates 25% more variation between the observed and the model responses.

The Rasch model is probabilistic or stochastic. Given this, we expect that the data show such behaviors as being “closer to the Guttman-style response string” which indicates that there is “less variation than modeled” in the data set (underfit). It is also possible to find more variations than modeled, which is an indicator of a more haphazard response string (overfit). These two cases imply different measuring problems. Underfit implies the inability of the “ensuing measures” (Bond & Fox, 200, p. 240). On the other hand, overfit is misleading in that it over-estimates the quality of the measure, although the problem it creates will be usually constraint to inflated reliability and smaller standard errors of measurement. Thus, items that “[s]tatistically and substantially” do not add any information to measure the trait. Bond and Fox further propose that overfitting items may
be locally dependent. This is a dilemma which must be considered to decide on the fit of an item, infit or outfit mean square (MNSQ).

According to Linacre (2008), the analyst can use either which is more “useful” for their analysis. Whereas the outfit mean square is “sensitive to grossly unexpected response”, infit mean square is more sensitive to outliers (p. 245). Standardized $Z$ (ZSTD) “is only useful to salvage non-significant MNSQ $> 1.5$, when sample size is small or test length is short” (Ben Wright; as cited in Linacre, 2008, p. 265). All in all, high infit MNSQ statistics indicate that the items are mal-functioning for the testees. This threat is more serious than significantly underfitting outfit MNSQ indexes, which indicate the existence of a few random responses by low-performers. Thus, infit MNSQ is considered more informative. In this study, all infit MNSQ values fall within the acceptable range. Nevertheless, some items cause some concerns according to their outfit MNSQ indexes (8, 9, 14, and 19). As stated above, Q8 and Q9 are the easiest items, so their overfit is predictable. Q14 was difficult which was argued to be mainly ascribable to its format. Q19 which measures the ability to understand trivial information is misfitting perhaps due its intent (see below). There is quite sufficient evidence for acceptable item/person fit validity of the data.

To further inspect this finding and also investigate the dimensionality of the test, a principal components analysis of residuals (PCAR) was run. Results from this analysis will help decide on the response validity of the data. The PCAR may be confused with the principal components analysis (PCA) in the Classical Testing Theory (CTT), but unlike the PCA which extracts factors, components observed in PCAR “show contrasts between opposing factors, not loading on one factor” (Linacre, 2005, p. 216; emphasis in
original). PCAR is similar to the unrotated factor matrix in PCA. As Linacre (2005) states, if any factor extracts more than 3 units, it is likely that we have come across a second dimension. According to Hambleton, Murray, and Williams (1983) “Test items within a content category may show a different pattern of residuals if they “tap” a different trait from the one measured by the items in the other content categories” (p. 12).

In this analysis, we need the total number of units of randomness which equals the number of measured variables. In the present study this index is 40 since there are 40 items or measured variables. In other words, based on the Rasch model, each variable contributes one unit of randomness. Table 3 summarizes the findings of the PCAR. Five factors (contrasts) were observed from the PCAR (Table 3). Factor 1 extracts only 2.5 units (1.6%) out of 40 units of variable residual variance noise. This denotes that the contrast between the strongly positively loading items and the strongly negatively loading items on the first factor in the residuals has the strength of about 3 items.

Table 3

<table>
<thead>
<tr>
<th>Index</th>
<th>Empirical</th>
<th>Modeled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total variance in observations</td>
<td>161.4</td>
<td>100.0%</td>
</tr>
<tr>
<td>Variance explained by measures</td>
<td>121.4</td>
<td>75.2%</td>
</tr>
<tr>
<td>Unexplained variance (total)</td>
<td>40.0</td>
<td>24.8%</td>
</tr>
<tr>
<td>Unexplained variance explained by 1st factor</td>
<td>2.5</td>
<td>1.6%</td>
</tr>
<tr>
<td>Unexplained variance explained by 2nd factor</td>
<td>2.4</td>
<td>1.5%</td>
</tr>
<tr>
<td>Unexplained variance explained by 3rd factor</td>
<td>2.1</td>
<td>1.3%</td>
</tr>
<tr>
<td>Unexplained variance explained by 4th factor</td>
<td>2.0</td>
<td>1.3%</td>
</tr>
<tr>
<td>Unexplained variance explained by 5th factor</td>
<td>1.9</td>
<td>1.2%</td>
</tr>
</tbody>
</table>

According to Linacre (2005, 2008), this contrast is too small to be considered as even a “conspicuous branch.” That is, the Rasch dimension dominates almost 47 times the second dimension and the second dimension is not
noticeable at all. Although this contrast (dimension) comprises almost 3 (2.5) items, it does not have substance to be considered as a dimension beside the Rasch dimension. By the same token, other contrasts comprise a tiny portion of the variance. To further support the decision that this dimension is not considerable enough to merit a separate dimension, the eigenvalue graph of the items was investigated. Figure 4 shows that four items encapsulated in a rectangle have higher loadings on the contrast. However, analyzing these items did not reveal any common pattern among them.

*Figure 4*
The eigenvalue graph of the items loading on the biggest contrast

Therefore, we can state with certitude that the test scores enjoy a considerable unidimensionality and response validity. This conclusion is further borne out by the high variance explained by the measures. It has been expected by the Rasch model that the measures will explain 77% of the variance, and the empirical observation is very close to this observation (75.2%). This variance, according to Linacre (2005, 2008) must exceed 50% in order to be considered significant in the measurement.
At this stage of the study, it is in place to examine the local independence of measured variables. Conventionally, unidimensionality and local independence are two criteria that must be established in the Rasch analysis. To analyze this question, the correlation statistics between standardized residuals were analyzed (Linacre, 2005, 2008). Table 4 summarizes the results from this analysis. The correlations which are displayed are all weak. Thus, items are certainly locally independent, denoting that answering or missing a questions will not affect answering or missing another question in the test. To recapitulate briefly, in this section evidence was presented to indicate the presence of construct under-representation, yet the test scores did not show symptom of sever construct-irrelevant factors although construct-irrelevancy will be further investigated. Overall, all validity classes but meaning validity (from item calibration) were sufficiently established.

Table 4
Largest standardized residual correlations used to identify dependent items

<table>
<thead>
<tr>
<th>Residual Correlation</th>
<th>Item</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.32</td>
<td>Q38</td>
<td>Q40</td>
</tr>
<tr>
<td>0.32</td>
<td>Q33</td>
<td>Q34</td>
</tr>
<tr>
<td>0.28</td>
<td>Q38</td>
<td>Q39</td>
</tr>
<tr>
<td>0.28</td>
<td>Q16</td>
<td>Q19</td>
</tr>
<tr>
<td>.27</td>
<td>Q17</td>
<td>Q32</td>
</tr>
<tr>
<td>-0.33</td>
<td>Q6</td>
<td>Q37</td>
</tr>
<tr>
<td>-0.31</td>
<td>Q21</td>
<td>Q40</td>
</tr>
<tr>
<td>-0.31</td>
<td>Q2</td>
<td>Q38</td>
</tr>
<tr>
<td>-0.27</td>
<td>Q2</td>
<td>Q40</td>
</tr>
<tr>
<td>-0.27</td>
<td>Q18</td>
<td>Q27</td>
</tr>
</tbody>
</table>

6.2. Differential Item Functioning

According to Bond and Fox (2007) “when an item’s difficulty estimate location varies across samples by more than the modeled error, then prima facie evidence of DIF exists”
(p. 92). DIF analysis is also a “useful way of examining tests for their fairness” (Geranpayeh & Kunnan, 2007). Geranpayeh and Kunnan point out that “a large DIF value could mean the test item is measuring additional dimensions differently across the reference and the focal groups” (p. 191).

WINSTEPS version 3.57 was used to invoke DIF measures for the sub-groups and also to compare them via t-tests. Linacre (2005) indicates that this “[t]-test is a two-sided test for the difference between two means (i.e., the [DIF] estimates) based on the standard error of the means (i.e., the standard error of the estimates). The null hypothesis is that the two estimates are the same, except for measurement error” (p. 223). In addition, standardized errors of measurement and also degrees of freedom are invoked from WINSTEPS. For space consideration, Table 5 demonstrates only the items that have significant DIF among different education groups.

In Table 5, ‘class’ denotes the category based on which the data was classified and analyzed (0: no university degree; 1: BA or BSc degree; 2: MA or MSc degree; 3: PhD). The first DIF column from left belongs to the first class (class A) and the fourth one to the second class (Class B). For example, according to Table 5, the local difficulty of Q5 for Class A is -0.62, whereas this index for Class B is 2.94. The difference between these two statistics exceeds 0.5, and is therefore flagged for further analysis (Linacre, 2008). Another column is DIF S.E., which is the Standard Error of the DIF. The column named Welch t “gives the DIF significance as a Welch (Student’s) t-statistics” (Linacre, 2008, p. 311). In this case, p is the probability of the reported t value with its degree of freedom (d.f.). Every DIF is assigned a standard error of measurement. The difference between DIF sizes is labeled DIF contrast which, according to Linacre (2005), “is a log-
odds estimate, equivalent to a Mantel-Haenszel DIF size” (p. 245). Mantel-Haenszel reports DIF tests for binary data. The $p$ value shows the significance of the differences observed. We usually need probability values to be very small if we are willing to find DIF among items, so that the DIF effect is not due to random accident. The significant statistical difference can be further supported by substantive difference.

Table 5

<table>
<thead>
<tr>
<th>Class</th>
<th>DIF</th>
<th>DIF</th>
<th>Class</th>
<th>DIF</th>
<th>DIF</th>
<th>DIF</th>
<th>Welch t</th>
<th>d.f.</th>
<th>p</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-.62</td>
<td>.60</td>
<td>3</td>
<td>2.94</td>
<td>1.39</td>
<td>-3.56</td>
<td>-2.35</td>
<td>16</td>
<td>.0322</td>
<td>Q5</td>
</tr>
<tr>
<td></td>
<td>-.38</td>
<td>.37</td>
<td>2</td>
<td>1.39</td>
<td>1.39</td>
<td>-3.31</td>
<td>-2.29</td>
<td>40</td>
<td>.0271</td>
<td>Q5</td>
</tr>
<tr>
<td></td>
<td>.04</td>
<td>.37</td>
<td>2</td>
<td>1.39</td>
<td>1.39</td>
<td>-5.68</td>
<td>-3.53</td>
<td>40</td>
<td>.0028</td>
<td>Q14</td>
</tr>
<tr>
<td></td>
<td>-2.74</td>
<td>.80</td>
<td>3</td>
<td>2.94</td>
<td>1.39</td>
<td>-3.56</td>
<td>-2.35</td>
<td>16</td>
<td>.0322</td>
<td>Q5</td>
</tr>
<tr>
<td></td>
<td>-1.60</td>
<td>.42</td>
<td>3</td>
<td>2.94</td>
<td>1.39</td>
<td>-3.31</td>
<td>-2.29</td>
<td>40</td>
<td>.0271</td>
<td>Q14</td>
</tr>
<tr>
<td></td>
<td>-1.66</td>
<td>.60</td>
<td>3</td>
<td>2.94</td>
<td>1.39</td>
<td>-4.54</td>
<td>-3.11</td>
<td>70</td>
<td>.0447</td>
<td>Q10</td>
</tr>
<tr>
<td></td>
<td>-1.35</td>
<td>.61</td>
<td>2</td>
<td>2.04</td>
<td>1.39</td>
<td>-4.60</td>
<td>-3.03</td>
<td>34</td>
<td>.034</td>
<td>Q14</td>
</tr>
<tr>
<td></td>
<td>-1.35</td>
<td>.61</td>
<td>2</td>
<td>2.04</td>
<td>1.39</td>
<td>-3.31</td>
<td>-2.29</td>
<td>40</td>
<td>.0028</td>
<td>Q14</td>
</tr>
<tr>
<td></td>
<td>-1.66</td>
<td>.42</td>
<td>2</td>
<td>2.04</td>
<td>1.39</td>
<td>-4.54</td>
<td>-3.03</td>
<td>34</td>
<td>.034</td>
<td>Q10</td>
</tr>
<tr>
<td></td>
<td>-1.66</td>
<td>.42</td>
<td>2</td>
<td>2.04</td>
<td>1.39</td>
<td>-2.52</td>
<td>-3.40</td>
<td>46</td>
<td>.0497</td>
<td>Q14</td>
</tr>
<tr>
<td></td>
<td>-1.66</td>
<td>.42</td>
<td>2</td>
<td>2.04</td>
<td>1.39</td>
<td>-3.31</td>
<td>-2.29</td>
<td>40</td>
<td>.0028</td>
<td>Q14</td>
</tr>
<tr>
<td></td>
<td>.51</td>
<td>.65</td>
<td>2</td>
<td>2.18</td>
<td>4.1</td>
<td>4.54</td>
<td>3.03</td>
<td>70</td>
<td>.0004</td>
<td>Q19</td>
</tr>
<tr>
<td></td>
<td>2.11</td>
<td>.61</td>
<td>2</td>
<td>2.18</td>
<td>4.1</td>
<td>4.54</td>
<td>3.03</td>
<td>70</td>
<td>.0004</td>
<td>Q19</td>
</tr>
<tr>
<td></td>
<td>2.11</td>
<td>.61</td>
<td>2</td>
<td>2.18</td>
<td>4.1</td>
<td>4.54</td>
<td>3.03</td>
<td>70</td>
<td>.0004</td>
<td>Q19</td>
</tr>
<tr>
<td></td>
<td>2.11</td>
<td>.61</td>
<td>2</td>
<td>2.18</td>
<td>4.1</td>
<td>4.54</td>
<td>3.03</td>
<td>70</td>
<td>.0004</td>
<td>Q19</td>
</tr>
<tr>
<td></td>
<td>2.11</td>
<td>.61</td>
<td>2</td>
<td>2.18</td>
<td>4.1</td>
<td>4.54</td>
<td>3.03</td>
<td>70</td>
<td>.0004</td>
<td>Q19</td>
</tr>
</tbody>
</table>

As a rule of thumb, the bigger the sample size, the more accurate the DIF analysis. Linacre (2005) states that “There is no minimum size, but the smaller the group size, the less sensitive the DIF test is statistically. Generally, results produced by groups sizes of less than 30 are too much influenced be idiosyncratic behavior to be considered dependable” (p. 245). According to Table 5, some items have functioned differently across groups with different educational backgrounds. These items are Q5, Q10, Q14, Q16, Q19, Q22, Q23, and Q35. Flagging these items at this stage helps analyzing their content. Content analysis is the successive stage of DIF which is carried out by experts or the researcher (Geranpayeh & Kunnan, 2007). The most useful procedure is to ask the item writers of the test to perform the task.
Table 6

*Rasch DIF for not/taking IELTS preparation courses (only significant observations are reported)*

<table>
<thead>
<tr>
<th>Class</th>
<th>DIF</th>
<th>S.E.</th>
<th>Class</th>
<th>DIF</th>
<th>S.E.</th>
<th>DIF</th>
<th>S.E.</th>
<th>d.f.</th>
<th>p</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>A with</td>
<td>-1.58</td>
<td>.65</td>
<td>B without</td>
<td>-3.28</td>
<td>.49</td>
<td>1.70</td>
<td>.49</td>
<td>95</td>
<td>.0394</td>
<td>Q8</td>
</tr>
<tr>
<td>A with</td>
<td>-0.89</td>
<td>.54</td>
<td>B without</td>
<td>-3.54</td>
<td>.54</td>
<td>2.65</td>
<td>.54</td>
<td>95</td>
<td>.0008</td>
<td>Q9</td>
</tr>
</tbody>
</table>

On the other hand, two items exhibit different functioning for the groups with/out taking IELTS preparation courses (Table 6). From the Rasch analysis of these two items (see section 4.4.), we learn that they belong in the group of very simple questions. This observation is interesting because the simplest items have shown different patterns for the two groups. So, Q8 is 1.70 logits and Q9 is 2.65 logits easier for the participants who have taken IELTS preparation courses.

Table 7

*Rasch DIF genders (only significant observations are reported)*

<table>
<thead>
<tr>
<th>Class</th>
<th>DIF</th>
<th>S.E.</th>
<th>Class</th>
<th>DIF</th>
<th>S.E.</th>
<th>DIF</th>
<th>S.E.</th>
<th>d.f.</th>
<th>p</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>-.57</td>
<td>.37</td>
<td>Female</td>
<td>.84</td>
<td>.32</td>
<td>-1.41</td>
<td>.32</td>
<td>97</td>
<td>.0051</td>
<td>Q20</td>
</tr>
<tr>
<td>Male</td>
<td>1.61</td>
<td>.44</td>
<td>Female</td>
<td>.33</td>
<td>.32</td>
<td>1.28</td>
<td>.32</td>
<td>97</td>
<td>.0209</td>
<td>Q23</td>
</tr>
</tbody>
</table>

As demonstrated in Table 7, the last DIF was run according to gender and two differently functioning items were found (Q20 and Q23). These items must be content-analyzed to find the possible reasons. Interestingly, significant DIF indexes were not observed in other classifications such as the nationality and age, which suggests that the operationalized construct has not behaved differently across different age and nationality groups.
7. Discussion

As reviewed, meaning validity of the items was not proved, yet other classes of Rasch validity were established. Accordingly, there was evidence for construct under-representation, yet not enough proof for construct irrelevancies. Further, unidimensionality and local independence as two indispensable criteria of the Rasch analysis were also observed. To further analyze the items, as DIF analysis was employed. Items which behaved differently across different groups were then scrutinized. In the content analysis of items and the passage, no clue was found to indicate that an item could be biased towards any gender in particular, yet the Rasch analysis showed that questions 20 and 21 functioned differently for two genders. Item 20 was easier for female test takers whereas item 21 was easier for male participants. Reviewing these two questions, it was found that they belong in the limited production (LP) item type. The answer to Q20 is golf club, which was written as ‘gulf club’ by some participants and missed by others. In this item, spelling appears problem-causing. However, it may be that the testees intended the correct response and as long as “the correct response was intended”, according to Hughes (1989), “there is no reason to deduct points for errors of grammar or spelling” (p. 139). In the same way, Q21 is also a LP item that requires a two-word answer (A plus/A+). Missing either would mean missing the item. Among wrong answers, some testees had written ‘A’ and some ‘plus’. Overall, this finding together with the observation mentioned above hints us to the presence of item format effect which has been detected by the Rasch DIF analysis. This is a piece of evidence indicating limited production format can cause some variance in the observed scores for a special group. Previous research has also proved that the length of required response
affects testees’s performance on listening tasks (Jensen, Hansen, Green, & Akey, 1997; Buck & Tatsuoka, 1998).

On the other hand, questions (5, 10, 14, 16, 19, 22, 23, and 35) were found to function differently across different educational levels. The common crux in all observed DIF values is that the higher educational groups have performed better on them. In other words, these questions are biased toward people with higher education levels. In this situation where invariance required for a sound measurement is not maintained, the theory-informed assessment requires further explanation why such phenomenon has been observed. Two explanations are offered here. This phenomenon can be due to the working memory overload. Short term memory is the active unit when processing the explicit and trivial information\(^3\) (van Dijk & Kintsch, 1983). So, detecting three words in the stream of a long discourse, keeping them in mind, and writing them while still being attentive to the rest of the aural message can create the ‘articulatory loop’ (Cook, 2001) and some part of the message may be lost soon. According to Coleman and Heap’s (1998) study on the IELTS Listening Test, this is a problem in the listening module of the test. This issue should be further investigated through retrospection studies.

Furthermore, reviewing Q5 which is a MCQ, it was found that to understand the item stem (and not the answer), the testee needs to do a ‘close paraphrase’. A close paraphrase happens when the listener/reader needs to paraphrase the sentence to comprehend it (Nissan, Devnicenzi, & Tang, 1996). Accordingly, if an item is intended to directly test whether a testee has comprehended, say, the price of fuel in 2008 or the

---
\(^3\) Analysing the underlying factors of the questions (e.g., whether they test understanding explicitly/implicitly stated information), which is performed through factorial studies, is not the purpose of this study; however, factorial studies can reveal invaluable facts about the underlying traits of the items and test structure (see Brown, 2006; Messick, 1994).
name of a product or person, this item is a question that assesses the ability to comprehend explicitly spoken information. But if the item contains sentences/phrases that are to be paraphrased (exact words are not heard), then a close paraphrase is needed to understand it. This should not be confused with inferencing (see Hildyard, & Olson, 1978). Only in this manner will the participant be able to answer the question correctly. This paraphrase is supposedly made in order for the testee to understand the stem and understanding the stem will help them choosing the best alternative. Item 5 and the tape script follow.

5. Which woman had once had her bag stolen in the past?

A. the elderly woman  B. Mrs Reynolds  C. Mrs Reynolds’ friend

Text: All in all, she was pretty calm. I don’t think I would have been, but apparently it had happened to her once before so maybe that’s why.

The answer is underlined. In order for the listener to answer the question, it is necessary to paraphrase the stem into a sentence like the underlined sentence. In other words, the answer to the question is not found if the listener only relies on their understanding of the arrangement of the words in the audio input. They must go further and apply some modifications to the structure of the sentence to match it with the audio input so that they will understand and answer it. Of course, according to Nissan et al. (1996) the concept of close paraphrase can cause problem at times for the researcher. It may not be really evident if the listener has to get involved in close paraphrasing or another sort of processing like inferencing.
The same pattern is also found in Q23. Thus, if the demands of the question in its stem are not apprehended via doing a close paraphrase then in most likelihood testees may miss the items even if they have comprehended the audio inputs. This is a construct-irrelevant factor\(^4\) (Buck, 2001). However, Q10 appears tricky in the sense that many students had written ‘small scarf’ in lieu of small scar, which is the answer. It is possible that the word ‘scar’ has not been familiar to the lower education groups. Field (2004) argues that this phenomenon is observed under some circumstances such as “the constraining influence of the lexical set”, substituting bottom-up processing with top-down processing, and substituting a “word phonologically similar to the target” (p. 372). Top-down processing proceeds from the schemata stored in the brain to corroborate or disprove the hypotheses that the listener has made. But bottom-up processing is characterized by a data-driven nature. It is initiated and directed by the audio message and is built upon processing and comprehending words and structures. More specifically, Field classifies the reactions of testees when they are confronted by unfamiliar lexis into three categories:

1) They try to find a phonological expedient, such as ‘assay’ for ‘essay’, or ‘scarf’ for ‘scar’, in the current study.

2) They give up when they cannot find such a match and as a result they submit a “blank response.”

3) They try to find a lexical match which is extremely similar to the unknown word, such as ‘know the sex’ for ‘note the sex’, in the present study.

\(^4\) We should note that fit indexes are indicators of the presence/absence of construct-irrelevant factors (see Bond, 2003). In the present study, whereas fit indexes showed that only items 14 and 19 may be misfitting (according to their outfit MNSQ), this was the DIF analysis that revealed further facts about construct-irrelevancies.
In Item 10 and similar ones such as 20, 21, 29, etc., attempt 1 was not very much observed, but it seems that attempts 2 and 3 prevail. This researcher agrees with Field (2004) that quantifying these observations will not portray the significance of this observation. Let’s suffice it that

… there is evidence here for a strategy which is neither bottom-up nor top-down but is lexical – a rough attempt at a one-to-one match with a known item which potentially overrules contextual information and modifies perceptual. This is a type of process that is often overlooked in the ‘bottom-up’/‘top-down’ controversy…. it qualifies as ‘top down’, but not in the way in which the term is usually interpreted…. (p. 373)

This strategy is seemingly compensatory; not to discount Field’s (2004) proposal suggesting that it best matches top-down processing, this researcher believes that most of the wrong answers in the present study can be manifestations of bottom-up processes which are dominant when a respondent answers items testing comprehension in explicit items. The difference between Field’s study and the present one lies mainly in the context. In Field’s study, context leads the low-skill testees towards top-down process by offering some contextual and perceptual cues, yet in the current study since there is no match between the wrong answers and the context, it is evidential that participants have shifted toward a bottom-up processing, to match what they know with the new lexis they hear without resorting to inferencing. But inferencing has obviously been the leading factor in Field’s study.

The DIF and content/response analyses further showed that the onset of words is robust for the groups in the study. The younger participants were found to create or
choose some codas that shaped the entire word as a known word that they were willing to substitute for the answer. As an example, whereas the correct answer to Q29 is ‘essay plan’, some participants had written ‘plain.’ Such an observation may simply testify to a tendency to substitute part of a word when there is a “perceptually similar alternative” (Field, 2004, p. 374). It cannot be due to simple erratic typography because its frequency was considerable and participants are assumed to be able to spell ‘plan.’ As will be further discussed below, these questions were found to be typically biasing toward participants with a higher level of educations. In this regard, it is argued that these respondents may have had a better command of English lexicon and as a result they did not have to go through any of the three processes articulated by Field (2004) to identify the word. But this speculation needs further inspection in future studies.

Q14 is a MCQ which is written together with Q15. Test takers are required to select two options out of the five options as the answers to Q14 and Q15, which are brought below.

Questions 14 and 15

Choose TWO letters, A-E.

Which TWO business facilities are mentioned?

A Internet access   B Mobile phone hire   C Audio-visual facilities   D Airport transport   E Translation services

Although the answers are explicitly stated, it seems that the joint length of these two items has affected the lower ability students’ performance. This observation is very similar to Coleman and Heap’s (1998) conclusion. Interviewing IELTS test takers, these
researchers found that when two items are located in the same place, this will “cause many problems” for the respondent. In Q14, therefore, this observation can be viewed as a sign of construct-irrelevancy.

Interestingly, Q16 resembles Q19 in format, i.e., both of them are LP types. These items are intended to measure students’ understanding of numbers, which is according to Shohamy and Inbar (1991) an item to assess the ability to comprehend trivial information. In Q16 where the correct answer is ‘35’ a frequently repeated observation was a wrong answer which includes ‘75’. This can indicate that answering limited production questions can at times cause problems. Furthermore, Shohamy and Inbar (1991) contend that the questions intended to measure trivial information “tend to make sever demands on the test takers’ memory load… [so] trivial questions should not be asked on LC [listening comprehension] tests” (p. 36). Together with Q22 and Q35, items 16 and 14 provide further evidence for our previous conclusion that LP questions behave differently from what the Rasch model explains, which has been previously found by Coleman and Heap (1998). They state that production questions in the IELTS Listening Test are the most difficult items for students. In a very similar study, Brindley and Slatyer (2002) also found that with limited production items, candidates “have about 30% less chance of being awarded the competency than if he or she were given [another item type]” (p. 380).

It was expected that the last section of the test would be the most difficult section due to its input and task properties. Research shows that input affects the test performance (Chaudron, Lubin, Sasaki, & Grigg, 1986; Chaudron & Richards, 1986; Chaudron, 1985; Bachman, 1990). The properties of an aural input can be its speed of
delivery, accent, pauses, grammatical structure, vocabulary density, etc. The Rasch DIF showed that the most difficult questions mainly were located in the last section of the test, irrespective of the exceptional items discussed. On the other hand, questions in section 1 were found the easiest. This finding is in conformity with Freedle and Kostin’s (1996) research who concluded when the passage was not academic, the items following it were easier.

As a summary, the following results were found in achieving the objectives of the study:

- The IELTS test scores showed evidence for all classes of Rasch validity but meaning validity. Meaning validity is evidence for efficient construct representation.
- Multiple choice questions (MCQ) are better item formats in testing listening since they do not pose construct-irrelevancies in measurement. They only require recognition and testees do not have to involve in production (Berne, 1993).
- MCQ, however, can cause DIF in listening test performance if the stem is too lengthy.
- MCQ can also cause DIF in listening test performance when understanding their stem needs doing close paraphrases (especially in a listening test like IELTS where candidates may develop an ‘articulatory loop’ (Cook, 2001) when answering questions). “Explicitly cued” items are easier (Brindley & Slatyer, 2002) than the items which involve participants in paraphrasing (Freedle & Kostin, 1996).
• In testing listening, production questions are not effective tools to measure the construct due to the limitations they pose on the listeners (see Hsiao-fang, 2004).

• Production items can be even more problematic when their answers comprise less common words that sound similar to more common words.

• There is evidence for lexical processing in this study, which may be confused with top-down processing; most probably, it is more of a bottom-up compensatory strategy in the present study although there is room for doubt.

• The IELTS Listening Test data was showed to be unidimensional and items were locally independent.

• This IELTS Listening Test showed some the symptoms of construct-irrelevant factors and construct under-representation.

• The Rasch analysis provides a utilitarian model to analyze unidimensionality (through principal component analysis of residuals), construct-irrelevant/under-representation factors (through fit indexes and the Wright map) in language testing.

• The Rasch DIF together with content analysis of items and passages is a fruitful way to investigate construct-irrelevancies.

8. Conclusion

It was observed that in many limited-production items, candidates encountered the problem of writing full answers so as to be considered a correct response. This is a problematic area in assessing second language listening, as Brown, Anderson, Shadbolt, and Lynch (1985) have also shown. According to Brown et al. (1985), it is possible that
candidates resort to over-use of their short term memory (STM) in writing the answers in a test of listening. Alternatively, since the IELTS Listening questions require no more than three words, the candidates might not have had sufficient writing skills to convert what they have heard to a three-word phrase. What Brown et al. articulate as causing variance in the listening tests are also observed in the current study and Coleman and Heap’s (1998) research. Put another way, since the students’ performance on LP items is generally weaker than that on other types, we find an effect of question format on their performance. Shohamy and Inbar (1991) also show that question types affect students’ performance.

There was strong evidence provided by the Rasch indicating that the test was to a great extent unidimensional. Although it was argued above that some factors such as the question types can cause some variance, it does not imply that the data set is not unidimensional. It is worth noting that no data set can be claimed as perfectly unidimensional. Therefore, a question such as ‘is my data set unidimensional?’ is not of interest in psychometrics (Linacre, 2005). A better inquiry would be “Is the lack of unidimensionality in my data sufficiently large to threaten the validity of my results?” (Linacre, 2005, p. 240). In other words, we can never arrive at a perfect unidimensionality although approximate measures are yielded in many analyses.

The Rasch analysis revealed the construct under-representation in the test. For the cohort taking the test and also for another sample size with a larger spread of ability levels there are not sufficient representative items available on the test. The test scores show some symptoms of construct-irrelevant factors (misfit). This study further showed that Wright and Stone’s (1999) fit/order validity framework is useful but it needs another
apparatus as a necessary extension. This apparatus in language testing and validation studies is DIF analysis jointly used with the analysis of test items and students responses. In this manner, the researcher/s can probe the issue of construct-irrelevancies more profoundly. Evidence to disconfirm the validity claim of the test score interpretations must be arranged against the evidence to corroborate it to make a decision. Although all validity classes but meaning validity were established in the present study, many items were showed to be involving construct-irrelevancies. This conclusion is also confirmed by research hitherto-undertaken. Therefore, this study does not yield sufficient evidence for establishing the construct validity of the test although it has provided some. But I do suggest carrying out more studies to prove or disprove the results from this study, because validation is not a static, one-stage analysis (Bachman, 2005).

9. Study limitations

This pilot study is limited in its sample size. So, as stated above, the results should not be considered definitive. More research is needed to dis/confirm the results from this study and expand the research line which has been stated in the present study. Also, results from this study may not be necessarily reflecting on the actual IELTS Listening Test issues.
References


Appendix 1

The consent form used in the study

Dear student,

I would like to invite you to participate in my PhD research. For my study, I will be collecting data from the listening section of the International English Language Testing System (IELTS) test. There will be a short training lesson prior to the main study. In this lesson, I will give you instructions on how to do the test. You will also practice an abridged IELTS test containing 10 questions. After a break, a test of IELTS Listening will be given to you. It lasts 40 minutes and contains 40 questions. While you are listening to the audio text, you are required to answer the questions. All of your answers are anonymous and confidential. You will also be asked to fill in a form to supply us with some demographic information. Below is a summary of the study plan.

1) Training and practice, 15 minutes;
2) A short break, 5 to 10 minutes;
3) Listening test, 40 minutes;
4) A questionnaire, 5 minutes.

I would appreciate that all parts of the test are completed. You can withdraw from the study at any time without penalty. There are no risks involved and the results will have no effect on your academic and educational status. When the project is finished, I will be happy to send you a summary of my findings from this study. If you agree to participate in the study, please sign the consent form below to return to me.
Best regards,
S. Vahid Aryadoust
English Language and Literature Department
National Institute of Education
Singapore
Tel: + 65 67903578
Email: ng0666907m@stdmail.nie.edu.sg

I would appreciate that all parts of the study are completed. However, you can withdraw from the study at any time without penalty.

I,_____________________(full name of the participant), have read the above comments and consent to voluntarily participate in the data collection session which comprises a training lesson followed by taking an IELTS Listening Test and a questionnaire.

Name:__________________________
Signature:_______________________
Email address:___________________
Date:___________________________

(This page will be enclosed to the letter of consent on the previous page.)
Appendix 2

Questionnaire to collect participants’ demographic information

Please Check √ or Write in the space provided.

1. Gender: Male □ Female □
2. I am [ _____ ] years old.
3. Nationality: [ _______________ ]
4. First Language: [ _______________ ]
5. Other languages spoken: [ ________________________________________ ]
6. Education: [ ___________________ ]

7. Which English tests have you taken before?
   A. TOEFL: ● iBT ● Computer-Based ● Paper-Based
   B. IELTS
   C. GRE

8. Have you taken ANY test Preparation courses before? How long?
   A. TOEFL _____ Time_______
   B. IELTS _____ Time_______
   C. GRE _____ Time_______

9. How do you evaluate the test event in this study?
   A. place Very good □ Good □ No idea □ Bad □ Very bad □
   B. sound system Very good □ Good □ No idea □ Bad □ Very bad □
   C. light Very good □ Good □ No idea □ Bad □ Very bad □
   D. temperature Very good □ Good □ No idea □ Bad □ Very bad □

Please write your comments (if any).
10. How do you evaluate the training lesson?
   A. *Explanations*  
      Very good □  Good □  No idea □  Bad □  Very bad □
   B. *The training test*  
      Very good □  Good □  No idea □  Bad □  Very bad □

Please write your comments (if any).