Macrothink Institute™

International Journal of English Language Education
ISSN 2325-0887
2013, Vol. 2, No. 1

A Critical Review of the Revised IELTS Speaking Test

Saeed Roshan

PhD student in languages

AUT Tower, corner of Wakefield and Rutland Streets

The Auckland University of Technology, Auckland, New Zealand

E-mail: saeedrosh10@gmail.com

Received: December 13, 2013 Accepted: December 26, 2013 Published: December 26, 2013

doi:10.5296/ijele.v2i1.4840 URL: http://dx.doi.org/10.5296/ijele.v2i1.4840

Abstract

The International English Language Test System (IELTS) is currently one of the English tests of repute, which is employed to assess the language proficiency of candidates planning to study or work in contexts where English is employed as the language of communication. This study is a critical review of the Revised IELTS Speaking Test (RIST) in order to highlight the strengths and weaknesses of the revised version. The findings indicate that the reduction from 5 phases to 3 phases in the structure, the introduction of an Interlocutor Frame (IF), the change of the rating system from holistic to analytic, and validity are the strong points of RIST. The weaknesses in the RIST could be subjectivity of the test, deviation from IF, and potential cultural bias. The study provides some recommendations for improvement of the Revised IELTS Speaking Test.

Keywords: IELTS, interlocutor frame, validity, reliability, subjective test

1. Introduction

The International English Language Testing System (IELTS) is devised to assess the language ability of candidates who would like to work or study where English is employed as a communicative language. Nowadays, IELTS is recognized as a prerequisite for English-medium study in higher education in most countries as well. Annually, more than 100,000 candidates participate in IELTS at 251 approved British Council and IDP Education Australia centres in over 105 countries. IELTS is employed to test all four language abilities of candidates. That is, Reading, Writing, Listening and Speaking. IELTS sets out to assess both academic and general English language proficiency of candidates (Taylor & Jones 2001).

One of the popular techniques for the assessment of oral language proficiency is the

conversational language proficiency interview. This kind of interview involves a face-to-face situation in which the interviewer queries a candidate on some specified topics (direct test). The belief behind this popular technique is that allows the interviewer a context in which to test the candidates' communicative and interactional skills (Brown 2000). The IELTS speaking test is one type of this test genre. Thus, it is regarded as a direct test.

The latest revision of the IELTS Speaking Test became operational in July 2001. The operational conditions of the Revised IELTS Speaking Test (***RIST***) involve one candidate and one examiner. The test embraces three phases which take between 11 and 15 minutes in total. In the first phase (4 to 6 minutes) candidates are asked to talk about themselves and their interests, and to reply to queries on familiar topics. In the second phase (3 to 4 minutes) the candidate must first talk on a topic presented on a cue card for approximately a minute and half; the remainder of the time is spent on preparation and on answering interviewer questions relating to the topic. In the last phase (4 to 5 minutes), the candidates are provided with an opportunity to discuss topics of a more abstract nature. These topics are thematically related to the second phase. For instance, if phase two concerns a holiday, then phase three can deal with travel or tourism (Issitt, 2008). The rating scale of *RIST* is from 0 to 9 (see appendix). Four analytic (sub) scales are employed to assess candidates' oral proficiency (performance), namely pronunciation, fluency and coherence, grammatical range and accuracy, and lexical resources (Taylor& Jones 2001).

This study deals with a critical review of ***RIST*** in order to highlight its strengths and weaknesses and provide some recommendations for improvement.

2. Strengths

2.1 Rationale for Revision

1. One of the advantages of the *RIST* is the reduction of structure from 5 phases to 3 phases. The original IELTS Speaking Test embraced 5 phases. However, the operational use of the original test indicated that phases 3 and 4 did not elicit the required performance from candidates; rather, they led to differing amounts and types of examiner-talk. *RIST*, on the other hand, has clear input and output requirements. Another advantage of the *RIST* is the design of the test; candidates move from familiar and less challenging topics to unfamiliar and more challenging ones (Taylor, 2001). In the first phase (introduction) examinees are asked to introduce themselves. This is a phase in which most candidates can easily engage, and in which their schematic knowledge can easily be activated. In the second phase (individual long turn), candidates can favour 1 minute planning time *prior to* explaining a topic specified on a cue card (*strategic planning* (Ellis 2003, 129)). However, there is some controversy concerning the advantages of planning times prior to a task Wigglesworth and Elder (2010) maintain that planning time possesses no positive advantages for candidates; however, it should be considered in the test development process in order to be fair and to increase the face validity of the test. However, Ellis (2003, p. 133) points out that an opportunity for strategic planning can have a positive effect on both fluency and complexity because it allow speakers to conceptualize what they want to communicate rather than how to say it. Weir, O'Sullivan and Horai (2006), further, point out that in test situation candidates

who have no planning time score considerably lower as opposed to candidates with one minute of planning time. Taylor (2001) emphasizes on the importance of preparation time in phase2 of *RIST* as it provides an opportunity for examiners to be free from their interactive role and focus on performance. This could be regarded as an important factor in maximizing the reliability and accuracy of the assessment. Taylor, further, points out that the topic on the cue card provides a "context and content points" to guide weaker candidates in particular. Phase 3 deals with a two-way discussion in operating conditions. The advantage of this phase is that it provides candidates with a more natural speaking situation. They engage in an interactive communication with less predictable questions which is very similar to what happens to candidates in real world situations. Thus, this could be regarded as the advantage of *RIST* in comparison to other tests such as TOEFL iBT in which candidates must interact with a computer.

2. The introduction of an Interlocutor Frame (IF) is the other significant change in the test procedure of the *RIST*. It is a script for the examiners' role and a guide to managing the test through phases 1 to 3. Thus, IF can be employed to enhance standardization of the test and reduce variability amongst different examiners (Taylor 2001).

3. The rating system of *RIST* has also changed from holistic to analytic rating. That is, rather than assessing candidates on their whole performance (holistic rating) examiners provide a separate assessment for each one of the four scales. Following this analytic rating the scores are combined into a single overall score (McNamara 2000, p. 44). Accordingly, analytic rating can enhance reliability due to more consistency in scores and also can reduce "rater-candidate interaction" (McNamara 2000, p. 99).

2.2 Validity

In two different studies, Brown (2006a) (2006b), seeks to manifest the validity of the analytic rating scales in the *RIST*. In the first study, Brown (2006a) applies an empirical analysis to reflect the validity of the analytic rating scales on the ground of candidates' discourse. The main aim of the study is to verify the use of descriptors to define the score points on the scales. Accordingly, Brown analyses the Speaking Test band descriptors and criteria key indicators in order to highlight relevant analytic categories for each of the 4 band scales (pronunciation, fluency and coherence, grammatical range and accuracy, and lexical resources). The data for the analysis are drawn from twenty IELTS Speaking test in various countries and with a range of proficiency levels. The findings indicate that although the study has some limitations on the grounds of scope, size and choice of analysis, overall the outcomes of this study support the validity of the Speaking Test band descriptors. Brown, further, states that "the overall tendency for most of the measures to display increases in the expected direction over the levels appears to confirm the relevance of the criteria they address to the assessment of the proficiency in the IELTS interview" (Brown 2006a).

In the second study, Brown (2006b) analyses the IELTS examiners' verbal reports and their responses to a subsequent questionnaire to investigate the validity (the interpretability and ease of application) of the analytic rating scales employed to assess candidates' performance in *RIST*. The evidence reflects a relatively good rating procedure. The examiners' reports

manifest their comfort and ease in employing the scales. Although examiners note some difficulties in making a clear distinction between scales and distinguishing levels, they report the consistency in their interpretations.

3. Weaknesses and Recommendations for Improvement

3.1 IELTS as a Subjective Test

1. It is difficult to discuss the drawbacks of an international English Test such as IELTS; however a review of the literature highlights a number of areas worthy of consideration. Bachman (1990, p. 76) distinguishes subjective test from objective test on the grounds of scoring procedure. In an objective test the candidate's response is corrected by "predetermined criteria" and examiners are not required to make a judgment. In a subjective test, such as an oral interview, the examiner must judge the correctness of the response in terms of his/her "subjective interpretation of the scoring criteria". If the examiner applies the same criteria and is consistent in his/her judgment of different candidates the result are more likely to be reliable. However, there is always the possibility in any rating situation that there will be inconsistency either in the rating criteria themselves or the way in which these criteria are applied. In order to examine test reliability, Bachman points out that examiners require to achieve at least two independent ratings for each individual speaking test sample (Bachman 1990, p. 179). Based on the above definitions *RIST* is a subjective test. With only one examiner there is a room for inconsistency within the individual ratings (intra-rater reliability). For instance, Read and Nation (2006) regard examiner inconsistency in rating lexical resources as a distinct component from the other three rating scales namely pronunciation, fluency and coherence, grammatical range and accuracy. Accordingly, they recommend "a revision of the rating descriptors for the lexical resource scale, so that they direct the examiners' attention to salient distinguishing features of the different bands".

2. As discussed above under "strength", *RIST*, now includes IF that is used by examiners to provide all candidates with the same test event. However, the subjective nature of *RIST* may occasionally lead to some deviations from IF and so compromise the validity of the test. For instance, as one part of their study, Seedhouse and Egbert (2006) analysed the transcripts of 137 audio-recorded test to illustrate several deviations from instructions. Their analysis reflected several deviations by examiners. They found, for instance, in some cases examiners aided candidates, the issue that leads to unfairness. The researchers therefore provide a number of recommendations in relation to examiner training, test design and instructions to ensure the validity of *RIST*. These include making sure that test variation and the length of preparation is acceptable. They also suggest providing trainee examiners with some examples from recorded data that reflect the failure of examiners to follow IF. These examples will clearly show examiners how they could compromise test validity. In another study, O'Sullivan and Lu (2006) examine the nature and location of examiners' deviations to see to what extent they impact examinees' oral performance. Their findings reveal that there are only a few deviations from IF in phase 1 and phase 2 of *RIST* and their impact on the language of the candidates is small. While examiners, in phase 3, sometimes show deviations from IF; the reason could be related the flexibility provided to examiners in this phase.

Another possibility might be related to the high cognitive load or lack of cultural or background knowledge inherent in the question types. Thus, the researchers recommend some flexibility in the IF so that examiners can paraphrase questions. However, although this deviation may have minimal impact on candidates' language, the researchers are unclear if it has any impact on their final scores.

3.2 Cultural Bias

On occasion examinees come across a number of topics and phrases which make it difficult for them to activate their schematic knowledge. Thus, they cannot actively engage in interaction to appropriately explain the topic. Therefore, candidates are likely to achieve a low score; whereas, they could be fruitful in another topic. Khan (2006) focuses on one of these drawbacks in the context of Bangladesh. Khan as an IELTS examiner provides a number of examples to reflect several cultural biases towards "Western culture and norms of behavior" in RIST content. Data were collected from questionnaires given to 18 examiners. Qualitative analysis manifested the presence of cultural biases inherent in topics, vocabulary, and terminology and question patterns of the speaking test. Khan argues that the presence of culturally unfamiliar features can confuse and stress candidates, leading to a negative impact on their oral performance. For instance, it is difficult for Bangladeshi IELTS candidates to respond to words such as such as "holiday" and "souvenir" as these words do not exist in their "linguistic and cultural repertoire". Tourism within Bangladesh is not strong as people lack the financial resources. Hence candidates possess no terminology to accompany such activities. Accordingly, Khan recommends that test designers of *RIST* consider the issue of cultural bias in the designing of tests; and possibly isolate topics and questions which disadvantage non-Western candidates. Seedhouse and Egbert (2006) also emphasise on the importance of culture issues. They recommend, for instance, the removal of "What shall I call you?" from questions because it results in significant problems. They point out that "the issue of how candidates and examiners address each other is a cultural one and may be adapted to the local conventions".

3.3 Financial Constrain and its Outcome on RIST's Reliability

Anyone who participates in *RIST* can easily see the concern of the examiners over the issue of time. As an examinee, I found it perplexing that my examiner regularly checked the time and occasionally interrupted me to change the topic of discussion. Ingram & Wylie (1993, cf. McNamara 2000 p. 102) point to a possible cause for this; in order to decrease the administration costs, the IELTS committee decided that the interview should be no more than 15 minutes. Financial factors also exclude a second examiner, a technique that could provide *RIST* greater reliability. Another important concern relating to *RIST* as a world test deals with the lack of control over the selection and the skills of the examiners. My examiner was an Indian woman with a strong Indian- English accent! These are the issues that the IELTS developer must pay more attention to in order to maximize the reliability of the test.

4. Conclusion

The revised IELTS Speaking Test is still one of the marked assessment systems for oral

proficiency. As discussed above, it has both strengths and weaknesses. The subjective nature of scoring is possibly one of the main problems of *RIST*. As McNamara (2000, p. 38) points out, the problem of subjectivity is an issue that should "be faced and managed". One way of managing this could be to introduce "*moderation meetings"* to prevent unfairness in the testing process (McNamara 2000 p. 44). It seems essential that examiners regularly participate in such meetings so that they can discuss the many challenges that they encounter during the test that can be discussed in ongoing moderating meeting. Increasing the time of the test to 20 minutes could also enhance both the naturalness and the reliability of the test by providing examiners with more opportunity to assess the examinees' oral performance ability. Ultimately, two examiners would also greatly increase the reliability of the test; however, an outcome of this would be a rise in administration fees. Instead, each center could choose one person as a supervisor who would randomly check the records of each examiner. Finally, it is seriously recommended that candidates' cultural issues be considered in test construction.

Reference

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

Brown, A, (2000). An investigation of the rating process in the IELTS oral interview. *Research Reports, 3,* 49-85. Retrieved from www.ielts.org/PDF/Vol3_Report3.pdf

Brown, A. (2006a). Candidate discourse in the revised IELTS Speaking Test. *IELTS research reports* (Vol. 6, pp. 1-19) IELTS Australia and British Council. Retrieved from https://www.ielts.org/pdf/Volume%206,%20Report%203.pdf

Brown, A. (2006b). An examination of the rating process in the revised IELTS Speaking Test. *IELTS research reports* (Vol. 6, pp. 1-30) IELTS Australia and British Council. Retrieved from www.ielts.org/PDF/Vol6_Report2.pdf

Ellis, R. (2003). *Task-based language learning and teaching.* Oxford: Oxford University Press.

Issit, S. (2008). Improving scores on the IELTS Speaking Test. *ELT Journal*, *62,* 131-138. http://dx.doi.org/10.1093/elt/ccl055

Khan, R. (2006). The IELTS Speaking Test: Analysing Culture Bias. *Malaysian Journal of ELT Research, 2*, 60-79. Retrieved from www.melta.org.my/modules/tinycontent/Dos/RubinaKhan.pdf

McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.

O'Sullivan, B., & Lu, Y. (2006). The impact on candidate language of examiner deviation from a set interlocutor frame in the IELTS Speaking Test. *IELTS research reports* (Vol. 6, pp. 1-27) IELTS Australia and British Council. Retrieved from www.ielts.org/PDF/Vol6_Report4.pdf

Read, J., & Nation, P. (2006). An investigation of the lexical dimension of the IELTS Speaking Test. *IELTS research reports* (Vol. 6, pp. 1-25) IELTS Australia and British Council.

Retrieved from www.ielts.org/PDF/Vol6_Report7.pdf

Seedhouse, P., & Egbert, M. (2006). The interactional organization of the IELTS Speaking Test. *IELTS research reports* (Vol. 6, pp. 1-45) IELTS Australia and British Council. Retrieved from www.ielts.org/PDF/Vol6_Report6.pdf

Taylor, L., & Jones, N. (2001). Revising the IELTS Speaking Test. *Research Notes*, *5*, 9-12. Retrieved from http://www.ielts.org/researchers/research/ielts_speaking_test.aspx/

Taylor, L. (2001). Revising the IELTS Speaking Test: developments in test format and task design. *Research Notes*, *5*, 3-5. Retrieved from http://www.ielts.org/researchers/research/ielts_speaking_test.aspx/

Weir, C., O'Sullivan, B., & Horai, T. (2006). Exploring difficulty in Speaking task: an intra –task perspective. *IELTS research reports* (Vol. 6, pp. 1-42) IELTS Australia and British Council. Retrieved from www.ielts.org/PDF/Vol6_Report5.pd

Wigglesworth, G., & Elder, C. (2010). An investigation of the effectiveness and validity of planning time in Speaking Test Task. *Language assessment quarterly*, *7*(1), 1-24. http://dx.doi.org/10.1080/15434300903031779

Appendix

Band 9 XPERT USER

Has fully operational commend of the language: appropriate, accurate and fluent with complete understanding.

Band 8 VERY GOOD USER

Has fully operational command of language with only occasional unsystematic inaccuracies and inappropriacies. Misunderstandings may occur in unfamiliar situations. Handles complex detailed argumentation well.

Band 7 GOOD USER

Has operational command of the language, though with occasional inaccuracies, inappropriacies and misunderstandings in some situations. Generally handles complex language well and understands detailed reasoning.

Band 6 COMPETENT USER

Has generally effective command of the language despite some inaccuracies, inappropriacies and misunderstandings. Can use and understand fairly complex language, particularly in

familiar situation.

Band 5 MODEST USER

Has partial command of the language, coping with overall meaning in most situations, though is likely to make many mistakes. Should be able to handle basic communication in own field.

Band 4 LIMITED USER

Basic competence is limited to familiar situations. Has frequent problem in understanding and expression. Is not able to use complex language.

Band 3 EXTERMELY LIMITED USER

Conveys and understands only general meaning in very familiar situations. Frequent breakdowns in communication occur.

Band 2 INTERNITTENT USER

No real communication is possible except for the most basic information using isolated words or short formulae in familiar situations and to meet immediate needs. Has great difficulty understanding spoken and written English.

Band 1 NON USER

Essentially has no ability to use the language beyond possibly a few isolated words.

Band 0 DID NOT ATTEMPT THE TEST

No assessable information provided.

www.ielts.org