SHIKEN: JALT Testing & Evaluation SIG Newsletter. October 2011. 15(2) 2- 9.

Application of the fusion model to while-listening performance tests

by Vahid Aryadoust (National Institute of Education, Singapore)

In the first installment of this article, I reviewed cognitive diagnostic assessment (CDA) and mentioned its advantages over other latent trait methods. I argued that the difficulty of a task can be accounted for by multiple factors or *attributes*[1]. Conventional unidimensional item response theory (IRT) models do not disseminate information concerning the factors attributing to task difficulty. On the other hand, the fusion model - which is a CDA model - partitions the difficulty parameter so as to furnish fine-grained information about the tasks and test takers' ability level. I further argued that granularity of the attributes is determined by researchers. In this installment the application of the fusion model to a while-listening performance (WLP) test of is described.

The term *while-listening performance (WLP) test* was first used by Aryadoust (in press) to refer to language listening tests where the test takers have to read and answer test items *while* they listen to oral input, and thus simultaneously engage in: (a) reading items; (b) listening to the oral input; (c) writing or choosing answers; and (d) following the oral input to move to the next item. The best-known WLP test is the International English Language Testing System (IELTS™).

"WLP tests represent the listening comprehension construct narrowly because they merely focus on pre-comprehension skills alongside the comprehension of details."

Aryadoust (in press) notes that the simultaneity of the cognitive and language production processes in WLP tests raises several important issues. First, these tests do not appear to distinguish listening comprehension skills from the ability to subsequently apply the comprehended input (i.e., supplying or choosing the answer) (Field, 2009), thereby misleading stakeholders about the listening construct and uses and interpretations of scores (Dunkel, Henning & Chuadron, 1993). WLP tests also seem to limit their focus to phoneme and word recognition sub-skills along with understanding details and surface information. For example, Geranpayeh and Taylor (2008, p. 3) described the IELTS listening test as a language assessment instrument developed "with some internal repetition," and test items as "focusing on explicit and easily accessible information". If this is true, then WLP tests represent the listening comprehension construct narrowly because they merely focus on pre-comprehension skills alongside the comprehension of details (see Shohamy & Inbar, 1991).

Furthermore, *multiple resources* theories of attention suggest that the capacity of working memory is limited (Wickens, 1976, 1980, 2007). Shifting attention rapidly from one modality to another in WLP tests can overload the memory span and adversely affect the test performance. This problem is particularly heightened for less experienced or low-ability test takers. For example, Eysenck and Keane, (1995) wrote that depending on task difficulty, practice, and similarity of tasks, attention can be divided among multiple tasks. Finally, Aryadoust (in press) adds:

> Given that WLP test takers' simultaneous exposure to oral and written inputs precludes note taking, it is likely that test takers who fall behind the stream of written/oral input miss some items not necessarily because of limited listening skills, but because of limited reading skills, memory span (Hildyard & Olson, 1978), test-taking strategies (Bachman, 1990), or test wiseness (Bachman, 1990; Kunnan, 1995), or because of other constraining influences (Field, 2009). (Aryadoust, in press, p. ##)

[1] Buck and Tatsuoka (1998) use the term *attribute* to refer to any factor that can affect test-taking processes. Attributes can be test taking strategies, item features, or cognitive and metacognitive strategies.

Due to the complexity of comprehension mechanisms in WLP tests, there are regrettably few studies investigating their structure. The present investigation seeks to describe the structure of the final section of the IELTS listening test (lecture comprehension) and provides a new window on some attributes affecting WLP lecture comprehension test performance. To serve this goal, I draw on empirical research into listening tests (e.g., Buck & Tatsuoka, 1998; Freedle & Kostin, 1999) as well as anecdotal or speculative taxonomies of listening comprehension sub-skills[2] (e.g., Richards, 1983) and propose a provisional attribute profile for the lecture comprehension section of the IELTS listening test. The profile is then subjected to fusion modeling.

Methodology

Participants

To select a sample similar to the actual candidates of the IELTS listening test, care was taken to invite those who either had recently taken the actual IELTS listening test or were enrolled in IELTS preparation courses at the time of the study. Participants were 209 multinational students with undergraduate and graduate degrees from China, Iran, Malaysia, and a few Arab states of the Persian Gulf. They were aged between 16 and 45 (M = 26; SD = 5.5). Eighty-nine participants were male and 120 were female. Informed consents were collected from participants and the invited individuals all opted to attend the study. Each individual was sent a report of his/her performance alongside suggestions to improve their listening skills. The study was conducted in the British Council in Malaysia and the National Institute of Education of Singapore.

Material

A version of the IELTS listening test selected from the *Official IELTS Practice Materials* (University of Cambridge Local Examinations Syndicate], 2007) was given to the participants. The test is the only available material endorsed by the developers and partners of the IELTS test. *Official IELTS Practice Materials* are created through the seven-stage Cambridge ESOL Question Paper Production Cycle process, the same process by which all Cambridge-developed tests are constructed. Previous IELTS studies including those funded by the University of Cambridge ESOL (English for Speakers of Other Languages) Examination Syndicate have selected commercially available IELTS practice materials given that "The IELTS partners do not release IELTS forms for research purposes" (Weir, Hawkey, Green, & Devi, 2009, p. 164). The Cambridge-developed materials, however, represent the actual IELTS tests; as Weir et al. put, such materials "[conform] to the IELTS specifications and [are] therefore representative of genuine IELTS test material" (Weir et al., 2009, p. 164).

The IELTS listening test consists of four sections: Sections 1 and 2 test the comprehension of everyday conversations, while sections 3 and 4 evaluate the comprehension of academic discourse. For the present study, I used section 4 - an academic talk with 10 test items in two types: sentence completion or gap filling (three items) and table completion (seven items). Performance of students on this section can be regarded as the representation of test takers' ability of lecture comprehension assessed by a WLP test (Field, 2009). Although with 10 items the probability of a non-Gaussian distribution becomes too large, the fusion model does not center around the normality assumption. (It is worth noting that as testified by skewness and kurtosis coefficients the univariate normality assumption in the present study was not violated).

[2] Since space limitations preclude a comprehensive exploration of the findings of these studies, readers are encouraged to consult the resources.

Data Analysis

Coding test items. I undertook a qualitative investigation of the test items, exploring the test items' structure and text content. For each test item, I noted a range of attributes including (a) the sub-skills tapped by the item, (b) task-related factors affecting participants' performance, and (c) text-related factors. The analysis was carried out twice with a one-week interval between to ascertain the intra-coder reliability. It is acknowledged that using two or more raters would offer greater reliability. However, given that experts familiar with the structure of the WLP tests were not available at the time of the study, it was decided to perform the coding twice with the same rater (the researcher) and control for the intra-reliability of the coding.

Fusion modelling. The fusion model (FM) is based around a Q-matrix, an incident matrix where the attributes are associated with test items. Findings of the item coding processes informed the development of the Q-matrix. It was found that a number of attributes would contribute to the successful performance on each test item.

I developed a Q-matrix of attributes in an endeavor to gauge their impact on test performance. The matrix was subjected to FM analysis. In the first round of the analysis, 11 attributed emerged. However, three attributes did not contribute to the statistical difficulty level of the test items as the initial fusion modelling indicated. Therefore, I removed the three attribute/item associations with r^*_{ik} estimates greater than .90 and respecified the matrix to obtain higher π^*_i indices (see Buck and Tatsuoka, 1998, for an account). For space reasons, I merely report on the findings of the finalized content list and FM analysis below.

I report three informative FM parameters: π^*_i, r^*_{ik}, and c_i, as they provide rich diagnostic input about each individual and test item and help evaluate the validity of the Q-matrix. Ideally, we would like to obtain r^*_{ik} estimates equal to or less than .90, which designates the power of the test items in discriminating masters from non-masters. If r^*_{ik} estimates fall below 0.50, the attribute is highly necessary to answer the question accurately (Roussos, Templin, & Hansen, 2005). Higher π^*_i estimates indicate that the attribute highly affects task difficulty. Finally, the parameter c_i is a "completeness index" and ranges from a low of 0 to a high of 3 (Montero, Monfils, Wang, Yen, & Julian, 2003). It will approach 3 if the attributes affecting the task difficulty are accurately and fully specified in the Q-matrix, and 0 if they are misspecified.

This investigation adheres to DiBello and Stout's (2008a) definition of "mastery probability" thresholds where .60 of probability to master an attribute is regarded as a *master* level, .40 is a *non-master*, and the area between .60 and .40 as the *indeterminate*[3] (*indifference*) region. Hence, test takers with a ≥.60 chance of accurately answering a test item are considered masters of that attribute and those with a ≤.40 chance are non-masters. Test takers with 40%-60% probabilities are not classified. The FM analysis was carried out on Arpeggio Suite for Windows, Version 3.1.001 (DiBello & Stout, 2008b).

Results

The Cronbach alpha internal consistency of the 10 test items was 0.813. This suggests a high degree of intercorrelation among the test items. The easiest item was Item 7 (M = 0.52; SD = 0.501) and the most difficult was Item 8 (M = 0.09; SD = 0.281) (see Appendix). Item coding generated a number of item-attribute associations which are displayed in Table 1.

[3] I am thankful to Tim Newfields for suggesting this term in lieu of the term *indifference*.

Table 1
Results of the Item Coding of the Ten Lecture Comprehension Items of the IELTS Listening Test

Attribute	Definition	Items associated with the attribute
1. Paraphrase	Listeners must keep the input in mind, read the test item and keep it in mind, and make a mental paraphrase of the aural message to match it with the written test item. For example, the text on the speed of a type of a bird says "*there is still some dispute about just how fast they can actually fly*". Item 2 reads: "*There is disagreement about their maximum*_____." The candidate must write (flight/flying) speed, synonymous to "*how fast they can actually fly*".	1, 2, 6, 7
2. Details	The ability of the listener to understand details such as names, specific pieces of information, and dates is tapped.	3, 4, 5, 8, 9, 10
3. Similar but misleading pieces of information	While listener is waiting for the right piece of information to arrive a few pieces of information that could fit the answer precede it, possibly confusing listeners. For example, the answer to Item 1 is *Australia*, which is a place name; the listener is awaiting a place name to pop up. But a few place names are heard before the answer, such as *South Pole* and the state of *Tasmania*. Given that test takers must make a spontaneous paraphrase of the aural stimuli to match it with the item and that they must keep a mental track of the place names that they hear, they may become distracted and miss the item.	1, 10
4. Paraphrasing the stem (synonyms)	To answer some items, candidates must understand synonyms. For example, the text related to Item 2 uses the word *dispute*, yet the item stem contains the word *disagreement*.	2
5. Accurate grammatical forms	Some test items require that the test taker recognize the exact grammatical points. For example, Item 3 requires a present participle: *"…the male spends some of his time*____." The answer is *looking or searching for food*. If *–ing* is dropped, the test taker might be penalized.	3
6. Low information density	That is, there is a relatively large amount of information not relevant to the answer in the text before arriving at the point where the answer lies. Answers do not appear rapidly in this sort of text.	1, 2, 3, 8
7. High information density	The answers to items 4 through 7 are condensed in one paragraph. High information density forces candidates to supply the answers more rapidly than items with less information density.	4, 5,6, 7, 9, 10
8. Repetition or paraphrase of the answer in the text	It seems that when information density is high, the answer to some - but not all - of the items is repeated or paraphrased in the text.	5, 9

There are eight attributes in Table 1 affecting item difficulty. It should be noted that initially 11 attributes emerged in the initial item coding, but due to their low effect on item difficulty, three (i.e., item format, number of words in the answer, and redundant information after the answer) were deleted. The actual effect of the attributes on the dependent variable (item difficulty) was explored through the FM, the results of which are presented in Table 2.

Table 2
Findings of the Analysis of the Finalized Q-Matrix on the Ten Lecture Comprehension Items of the IELTS Listening Test

Item	π_i^*	r^{*1}	r^{*2}	r^{*3}	r^{*4}	r^{*5}	r^{*6}	r^{*7}	r^{*8}	c_i
1	0.69	0.68	0	0.60	0	0	0.75	0	0	2.10
2	0.69	0.70	0	0	0.81	0	0.68	0	0	2.34
3	0.79	0	0.56	0	0	0.90	0.789	0	0	2.68
4	0.89	0	0.82	0	0	0	0	0.49	0	2.70
5	0.87	0	0.74	0	0	0	0	0.53	0.87	2.56
6	0.56	0.92	0	0	0	0	0	0.37	0	2.56
7	0.83	0.46	0	0	0	0	0	0.74	0	1.70
8	0.98	0	0.70	0	0	0	0.95	0	0	2.51
9	0.98	0	0.84	0	0	0	0	0.96	0.46	2.62
10	0.94	0	0.67	0.96	0	0	0	0.95	0	2.36

Reading across Table 2, the leftmost column gives the item number; the second column gives the π_i^* values. These range between 0.558 (Item 6) and 0.984 (Item 8). The relatively high π_i^* indices indicate that the postulated attributes account for task difficulty. The third through the tenth columns express the r_{ik}^* estimates. Most of these values are below 0.90 and only few exceed that, indicating the power of the test items to discriminate masters from non-masters. Some r_{ik}^* estimates (e.g., attribute 1 associated with Item 7) fall below 0.50, indicating that the attribute is highly needed to answer the item accurately. The completeness index c_i is given in the far right column and ranges between 1.707 (Item 7) and 2.700 (Item 4). This indicates that the attributes affecting the task difficulty are specified fairly accurately in the Q-matrix.

The number of respondents who are considered masters of the first through the eighth attributes is: 108, 114, 147, 141, 126, 118, 118, 152, 126, and 109; the remaining participants are classified as either nonmasters or else non-classified. The FM also provides classification information about the probability of mastery for each attribute. Table 3 gives this information about five randomly selected test takers. For example, the first test taker has likely mastered attributes 1 and 8, but is unlikely to have mastered attribute 6. Other attributes land on the borderline. Note that the test has been truncated; we can become more confident about the mastery of attributes on the borderline by giving the test takers a lengthier test. Table 3 also gives the observed and modeled scores. Comparing these sets of values is one of the measures of fit analysis in the FM.

Table 3
Classification Information of Eight Attributes Affecting Items Difficulty in the IELTS Listening Test

Student	Att. 1	Att. 2	Att. 3	Att. 4	Att. 5	Att. 6	Att. 7	Att. 8	Score	Modeled score
1	0.678	0.740	0.520	0.672	0.66	0.492	0.526	0.832	5	7.47
2	0.450	0.736	0.616	0.808	0.748	0.758	0.836	0.704	6	7.14
3	0.392	0.096	0.330	0.490	0.140	0.414	0.418	0.110	2	2.96
4	0.244	0.872	0.304	0.514	0.696	0.704	0.290	0.156	5	4.96
5	0.708	0.994	0.604	0.892	0.91	0.842	0.956	0.942	7	8.14

Note. Att. = Attribute. Names and definitions of attributes are given in Table 1.

To evaluate the fit of the model, I calculated the correlation between the estimated and modeled item *p*-values (item difficulty), which was 0.996 ($p < 0.001$). The significantly high correlation coefficient supports the fit of the model to the data. The computer program further gives a global measure of item fit which is the average difference between the observed and modeled *p*-values. In the present study, this index is 0.434 (0.826 - 0.392), which is a tolerable discrepancy. Roussos et al. (2005) argued that because the prime goal of the FM is to estimate the attribute mastery profiles of test takers and students, a slight discrepancy would not have a substantial influence over the results.

Discussion and Conclusion

This study set out to investigate the diagnostic features of section 4 of the IELTS listening test. IELTS is a while-listening performance (WLP) test where candidates must switch constantly across different modalities. I have used the fusion model (FM) to explore the variables that affect the difficulty of the test items.

The finalized item coding process generated eight attributes likely affecting item difficulty. Of these, attributes *paraphrase*, *details*, and *information density* (high/low information density) are common between the present and past studies (e.g., Buck & Tatsuoka, 1998), but the attributes *similar but misleading pieces of information*, *paraphrasing the stem (synonyms)*, *accurate grammatical forms*, and *repetition or paraphrase of the answer in the text* appear to be uniquely relating to WLP tests.

The FM aids in extracting the influential attributes taxing cognitive processes, though the decision on whether or not they are construct-irrelevant factors with high cognitive demands is left to the researcher. We can confidently argue that *paraphrasing the stem (synonyms)* and *accurate grammatical forms* are construct-irrelevant factors, because they are not directly related to the listening construct; they are germane to the execution of production skills. In addition, it is not unlikely that test takers understand a piece of information bearing the answer to a test item, yet when supplying the answer they become puzzled by virtue of either their lack of ability to paraphrase the stem or inaccurate application of grammatical knowledge. The latter case is testified by the presence of a number of incorrect answers in the participants' answer sheets. Such inaccurate answers would be considered accurate, had the answer key not required the candidate to supply the *exact* wording and/or to inflict a cognitively challenging grammatical transformation on the aural stimuli when writing it down.

"The [fusion model] aids in extracting influential attributes taxing cognitive processes, though the decision on whether or not they are construct-irrelevant factors with high cognitive demands is left to the researcher. We can confidently argue that paraphrasing the stem (synonyms) and accurate grammatical forms are construct-irrelevant factors."

This methodology has great potential for fair assessment, as it provides teachers with the opportunity to become aware of the underdeveloped attributes/skills of their students and offer them remedial actions. The model also helps language teachers and curriculum developers distinguish masters from nonmasters of each attribute and their proportion. This can inform the instruction, material design or selection, and even school policies. Kunnan and Jang (2009) argue that "diagnostic feedback can routinely be offered in all assessments, not just in so called diagnostic tests, and diagnostic feedback can reach its full potential of integrating assessment with teaching, learning, and the curriculum" (Kunnan & Jang, 2009, p. 622).

However, the current computer programs require fairly large samples, a requirement which might preclude using them for classroom assessment purposes. It is very desirable to see the new and less demanding (in terms of sample size) FM calibration techniques alongside other cognitive diagnostic assessment (CDA) methods and employ them at schools. I believe that this is not a far-fetched or quixotic undertaking, as a great number of school teachers in Hong Kong and relatively fewer in Singapore and China have now accommodated the use of such latent trait models as the Rasch model into their classroom assessments. Using the FM or other CDA models would seem to be equally viable in the context of in-house assessments.

References

Aryadoust, V. (in press). Differential item functioning in while-listening performance tests: The case of the IELTS listening test. *The International Journal of Listening*.

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

Buck, G., & Tatsuoka, L. (1998). Application of the rule-space procedure to language testing examining attributes of a free response listening test. *Language Testing, 15*, 119-157. doi: 10.1177/026553229801500201

DiBello, L., & Stout, W. (2008a). *Arpeggio documentation and analyst manual*. Chicago: Applied Informative Assessment Research Enterprises.

DiBello, L., & Stout, W. (2008b). Arpeggio Suite, Version 3.1.001, [Computer program]. Chicago: Applied Informative Assessment Research Enterprises.

Dunkel, P., Henning, G., & Chaudron, C. (1993). The assessment of an L2 listening comprehension construct: A tentative model for test specification and development. *Modern Language Journal, 77*, 180-191.

Eysenck, M. W., & Keane, M. T. (1990). *Cognitive psychology: A student's handbook*. Hove: Lawrence Erlbaum Associates Ltd.

Field, J. (2009). A cognitive validation of the lecture-listening component of the IELTS listening paper. In L. Taylor (Ed.), *IELTS research reports* (Vol. 9) (pp. 17-66). Canberra: IELTS Australia, Pty Ltd & the British Council.

Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing*, *16*(1), 2–32. doi: 10.1177/026553229901600102

Geranpayeh, A., & Taylor, L. (2008). Examining listening: Developments and issues in assessing second language listening. *Research Notes, 32*, 3-5.

Hildyard, A., & Olson, D. (1978). Memory and inference in the comprehension of oral and written discourse. *Discourse Processes, 1*(1), 91-107. doi: 10.1080/01638537809544431

Kunnan, A. J. (1995). *Test taker characteristics and test performance: A structural modeling approach.* Cambridge: Cambridge University Press.

Kunnan, A. J. & Jang, E. E. (2009). Diagnostic feedback in language assessment. In M. Long & C. Doughty (Eds.), *The handbook of language teaching* (pp. 610-625). Blackwell Publishing.

Montero, D. H., Monfils, L., Wang, J., Yen, W. M., & Julian, M. W. (2003, April). *Investigation of the application of cognitive diagnostic testing to an end-of-course high school examination.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL. Retrieved from http://www.education.umd.edu/EDMS/MARCES/mdarch/pdf/msde000005.pdf

Richards, J. C. (1983). Listening comprehension: Approach, design, procedure. *TESOL Quarterly, 17*, (2), 219-240. doi:10.2307/3586651

Roussos, L., Templin, J., & Hensen, R. (2005). *Theoretically grounded linking and equating for mastery/non-mastery skills diagnosis models.* Retrieved from http://www.measuredprogress.org/documents/10157/19213/GroundedLinking.pdf

Shohamy, E., & Inbar, O. (1991). Construct validation of listening comprehension tests: The effect of text and question type. *Language Testing, 8*(1), 23-40. doi: 10.1177/026553229100800103

University of Cambridge Local Examinations. (2007). *Official IELTS Practice Materials 2*. Cambridge: Author.

Weir, C., Hawkey, R., Green, A., & Devi, S. (2009). The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British University. In L. Taylor (Ed.), *IELTS research reports* (Vol. 9) (pp. 15-190). Canberra: IELTS Australia, Pty Ltd & the British Council.

Wickens, C. D. (1976). The effects of divided attention on information processing in manual tracking. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(1) 1-13. doi: 10.1037/0096-1523.2.1.1

Wickens, C. D. (1980). The structure of attentional resources. In R. Nickerson (Ed.), *Attention and performance* (VIII) (pp. 239-257). Hillsdale, NJ: Lawrence Erlbaum.

Wickens, C. D. (2007). Attention to attention and its applications: A concluding view. In A. F. Kramer, D. A. Wiegmann, & A. Kirlik (Eds.), Attention: From theory to practice. New York: Oxford University Press.

Appendix

Two sample test items from the IELTS test used in this study.

Questions 4 – 7

Complete the table below.

*Write **NO MORE THAN THREE WORDS** for each answer.*

Age of falcons	What occurs
[Items 4 through 6]	[Items 4 through 6]
1 - 12 months	More than half of the falcons **7** …….. [answer = ***die***]

Questions 8 – 10

Complete the notes below.

*Write **NO MORE THAN THREE WORDS** for each answer.*

Procedures used for field research on peregrine falcon chicks	
First:	Catch chicks
Second:	**8**……to legs