© Kamla-Raj 2016
PRINT: ISSN 0972-0073 ONLINE: 2456-6802
Anthropologist, 23(1,2): 218-225 (2016)
DOI: 10.31901/24566802.2016/23.1-2.27

A Tentative Research Mode of Comparability Study in Language Testing and Its Application

Wei Wang

Department of English, School of Humanities and Social Sciences, Xi'an Polytechnic University, No. 19, South Jinhua Rd., Beilin Dist., Xi'an City, 710048, P.R.China
E-mail: wang_wei0728@163.com

KEYWORDS Language Testing. Listing-Sub-Test. Research Mode

ABSTRACT A tentative research mode, based on the framework of CLA and TMF proposed by Bachman, was designed to guide comparability study in language testing in the hope that the mode proves to be reliable in this field. To achieve this, the paper compared two listening comprehension sub-tests, TEM4 and TOEFL, as to whether they measure a similar English listening proficiency and as to which one requires a higher level of communicative language ability. Experiment, questionnaire survey and test content analysis were conducted to seek answers to the research questions. Results indicate a relatively high complexity of TOEFL listening sub-test, which requires a higher level of communicative language ability. However, TEM4 listening sub-test is more reliable than its counterparts. Therefore, the present comparability study demonstrates that the research mode proves its reliability and validity, which is conducive to furthering language testing studies.

INTRODUCTION

A Tentative Research Mode

The majority of research in language testing is based on tests or questionnaires. Theoretically and practically, three elements and two instruments are usually involved in the field of language testing. The three elements are test takers, test scores and test papers while two instruments are tests and questionnaires. All these are quite essential especially when two tests are compared. To make the comparability study more convincing and scientific, the researchers should take all these into considerations. Based on Bachman's framework of Communicative Language Ability (CLA) and Test Method Facets (TMF) as well as on other researches (Bachman 1990; Yang 1998; Zhou 2004; Brunfaut 2015), a tentative research mode was proposed to see whether it is a reliable tool to conduct any research of this kind. To achieve this, a comparability study of English listening tests was conducted due to the fact that listening is one of the "under-researched aspect of assessment" (Harding et al. 2015) and that years of study can not assist second language learners in "comprehending proficient speakers in real-world settings" (Wagner 2015) (Fig. 1).

An Overview of TEM4 and TOEFL Listening Sub-test

Test for English Majors (TEM) is a test battery administered in China which consists of two tests—TEM4 and TEM8. TEM4 is designed for students majoring in English language and literature, and it is given near the end of the two years' foundation stage of a four-year degree program. The test is held every April by Shanghai International Studies University under the auspices of the Ministry of Education. As the only battery of tests targeting at English majors in China, TEM has already gained popularity nationwide and it is as a whole well established. As an indicator of the test taker's English proficiency and communicative language ability, the test can also be classified as criterion-referenced proficiency test.

TEM4 listening sub-test consists of dictation and listening comprehension which includes conversation, passage and news broadcast. The objective is to test candidates' ability to catch verbal messages. The test covers general topics related to daily life and matters related to study. The information of input is delivered at a speech rate of 120 wpm and read only once. There is a 5-second interval after each question item, which is printed on the test paper, as opposed to TOEFL PBT listening sub-test. Candidates are therefore required to select one best answer from the four options given.

Held by Education Testing Service (ETS), TOEFL is designed to measure the English proficiency of people whose native language is not English. It has been the leading academic English proficiency test in the world. The TOEFL test is, in nature, a norm-referenced proficiency

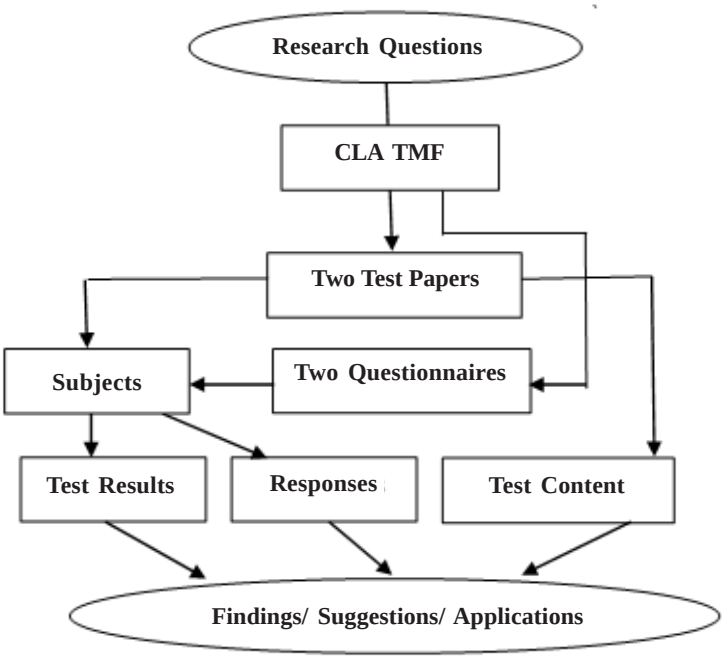

Fig. 1. Research mode in language testing
Source: Author

test (Bachman 1990), as opposed to TEM test. The test was and is offered in three formats: computer-based, paper-based and internet-based.

TOEFL listening sub-test (in paper-based format) measures ability to understand English as it is spoken in North America. The oral features of the language are stressed, and the problems tested include vocabulary and idiomatic expression as well as special grammatical constructions frequently used in spoken English. This section tests comprehension of main ideas, supporting ideas, important details, and inferences. The listening materials mainly focus attention on campus life and academic contents while the stimulus material and oral questions are recorded in standard North American English. The response options are printed in the test books. This section lasts 30 to 40 minutes and each question has a 12-second interval.

Rationale and Research Questions

TEM4 and TOEFL have been administered in China for years, and each draws a large population and has been considered an indicator of English language proficiency and a benchmark of further education. Both share something in common, and differences, of course, are in existence. The prominent factor directing the present research is, however, the statement that appears in the TEM4 Testing Syllabus, that is, "to comprehend the listening materials at the intermediate difficulty level (for example, the mini talk in TOEFL)". This makes one hold the assumption that the TEM4 listening sub-test measures the similar listening ability to that of TOEFL. It is therefore this assumption that initiates the present comparability study. Accordingly, three research questions emerge as follows:

1. Are the two tests similar in terms of testing English listening comprehension ability?
2. If yes, to what extent or in what aspects are they similar?
3. If no, to what extent or in what aspects are they different?

Objectives of the Paper

The primary purpose is to justify a proposed language testing research mode by conducting a comparability study of TEM4 and TOEFL listening sub-tests, and to elicit significant information conducive to research of its kind.

MATERIAL AND METHODS

In order to examine the assumption as well as to find out the answers to those research questions, experiment and test content analysis are conducted. Where the experiment is concerned, following research materials and methods are involved.

Instruments

The instruments adopted are two test papers and two questionnaires. The papers are authentic ones available. The reason why the once-administered authentic test papers are selected is that these test papers have been proved to be reliable and valid, and thus be representative. While TOEFL paper-based test has been out of use, it is still a good sample to conduct the comparability study.

Subjects

The subjects finally chosen are two different groups, a group of 92 non-English-major students majoring science and engineering, and another group of 140 English-major students. In a word, both groups take two different listening tests one after another within a certain period of time. When finishing tests, they are asked to complete a questionnaire survey.

Scoring

The scoring in this paper differs from that in many other studies, for the two tests employ a totally different scoring system: TEM4 is a criterion-referenced test while TOEFL is norm-referenced one. Except scoring the Dictation in TEM4, data collection is from a unique approach, counting the number of correct answers out of the total items. The eclectic approach could, to a certain extent, make the two test results comparable.

Item Facility

The approach of item facility is employed to elicit the information about which sub-test is relatively difficult for test takers to deal with. Since the approach can only be used to compute the item facility of objective test items, the following analysis just focuses on the listening comprehension with response-option items.

Item facility, also known as Facility value, is a measure of the ease of a test item (Richards 2002). It is the proportion of the students who answer the item correctly, and may be determined by the formula as follows: IF = R / N (R: number of correct answers; N: number of students taking the test).

Questionnaire Survey

The questionnaire for non-English-major group contains 14 questions and for English-major group 12. They are constructed on the basis of CLA and TMF framework, such as test taker's personal characteristics, topical knowledge, affective schemata and language of input and of the framework proposed by Rubin (1994).

Test Content Analysis

One of the first characteristics of a test that we examine is its content (Bachman 1990). Content analysis refers to a method in research used for analyzing and tabulating the frequency of occurrence of topics and other aspects of the content of written or spoken communication. According to Bachman (1990), when examining a test, we generally refer to a table of specifications and example items, or at least a listing of the content areas covered, and the number of items, or relative importance of each area. The consideration of test content is thus an important part of both test development and test use, as such, of a comparability study between two tests.

RESULTS

Test Result Analysis

Data was only collected from English-major group to make statistical analysis, for the population of this group is comparatively large enough, and their much exposure to English language makes them sensitive to even subtle differences and then lead to meaningful findings. The valid answer sheets for the objective items were brought from the 140 down to the 112, on the basis of which the test results were computed, compared and analyzed.

In Table 1, the ratio of Mean Score (MS) to Total Score is 67 percent for TEM4 and 54 percent for TOEFL, suggesting a relatively high difficulty of TOEFL in terms of objective multiple-choice items, and that the Standard Deviation (SD) is 6.5 for TEM4 and 7.28 for TOEFL, show-

Table 1: An overview of the test results

	TEM4				
	Dictation	*Conversation*	*Passage*	*News*	*Total*
Score	15	10	10	10	30
Mean	11.26	8	6.4	5.6	20
MS/TS	20/30 = 67%	Std. dev	6.5		
	TOEFL				
	Short Conversation	*Long Conversation*	*Short Talks*		*Total*
Score	30	9	11		50
Mean	16	6	5		27
MS/TS	27/50 = 54%	Std. dev	7.28		

ing that the score of TOEFL has a much higher degree of dispersion while that of TEM 4 a distribution with central tendency. To put it another way, higher SD index reflects a high discrimination power of a test.

Item Facility

The item facility of each item can be used to form the values of each section or part of the test, and of the whole test. Therefore, the author computed the item facility of each section of the two sub-tests respectively and then got the mean value.

Table 2 indicates that students who sat for TEM4 perform better than those for TOEFL on average, and that the TOEFL listening sub-test is more difficult than its counterpart (0.66>0.55), with the most demanding section, Short Talks, reaching the lowest index of 0.47.

Table 2: Comparison of item facility

TEM4	Conversation	Passage	News	Mean
	0.84	0.62	0.53	0.66
TOEFL	Short Conversation	Long Conversation	Short Talks	Mean
	0.53	0.65	0.47	0.55

To compute item facility of the subjective item, Dictation, another formula was adopted: P = MS/TS=11.26/15=0.75 (P = facility value; MS = mean score; TS = total score)

To examine the item facility of the entire TEM4 listening sub-test, a holistic approach is used.

Table 3 indicates that the TOEFL listening sub-test is more difficult than its counterpart in terms of test results analysis.

Table 3: Comparison of mean value of item facility

	Dictation	*Listening*	*Mean*
TEM4	0.75	0.66	0.71
TOEFL		0.55	0.55

Result 0.71 > 0.55; TOEFL > TEM4

Questionnaire Survey

A questionnaire was administered to 92 students majoring in science and engineering after they finished taking the two listening sub-tests. The responses are marked directly on the questionnaire for manual tabulation. The valid questionnaires were reduced to 70. Another questionnaire with slight revisions was administered to 140 English-major students with the same procedure as the previous one. The valid ones eventually decreased to 125.

Firstly, the responses from both groups are compared and analyzed. The attention is paid to the difficulty of the two listening sub-tests.

According to Chi-Square Tests of Independence, the statistical results for both groups, $\chi^2(23.6) > 7.82$, and $(24.86) > 7.82$, show that the subjects' attitude, in Table 4, towards the ques-

Table 4: The comparison of TEM4 and TOEFL listening subtests in terms of difficulty

Which test is more difficult?				
	TEM4	*TOEFL*	*Both*	*Unknown*
Non-English Group	31%	46%	13%	10%
English Group	32.8%	38.4%	18.4%	10.4%

df = k-1=4-1=3 α= 0.05
critical value = 7.82 (See Li 2001: 179)

tion is tendentious; namely, the results indicate a relatively high agreement in terms of test complexity that the TOEFL listening sub-test is more difficult than its counterpart and requires a relatively higher English listening proficiency.

Test Content Analysis

The description of the test content analysis consists of two parts: quantity of content input and task characteristics.

It should be pointed out that TEM4 contains a Dictation while the TOEFL does not, which may reflect the purpose or the different abilities to be measured respectively. Another difference is that the subjective and objective items account for a different percentile in the two tests respectively. Therefore, the part of Dictation is included when making the holistic analysis and comparisons while it is excluded when the numerical data and characteristics are needed.

Quantity of Content Input

According to Bachman (1997), input consists of the information contained in a given test task, to which the test taker is expected to respond. The analysis of content input can be carried out by such components as texts, stems and options in a test. This comparison needs a number of counting works, with qualitative analysis where necessary. The following two tables provide objective information on the quantity of input of two listening sub-tests respectively.

Tables 5 and 6 indicate some differences in quantity in each item type as well as in the whole length. TOEFL has a considerably large amount of input of words. Except one-turn-taking conversation, the amount of words in each text in TOEFL is also greater than that in TEM4. However, the following two data appear to be interesting. When taking TOEFL listening, test takers process more words per item than taking TEM4 listening (84.8>82). But while taking TEM4 listening, test takers processed more words per minute than taking TOEFL (164>141.3).

Task Characteristics

The above step focuses merely on the surface of the content input, that is, the amount of the input. Thus, the deeper nature of the task characteristics is essential to be examined through three steps: task type, sentence type and topical type.

The framework of task characteristics employed here is revised slightly from the Bachman's TMF theory and other researchers. According to Bachman, format is the presentation of task types, including channel, mode, vehicle, form, language and type.

Task type has to do with the way in which the input is presented. In Table 4, two of them

Table 5: Quantity of content input of TOEFL (Time: 30 min.)

Section I Part	*Stem No.*	*Stem Words*	*Text No.*	*Text Words*	*Option Words*	*Total Words*	*wpi*	*wpt*
A Short Conv.	30	168	30	1006	932	2106	70.2	33.5
B Longer Conv.	9	78	2	635	198	911	101.2	317.5
C Short Talks	11	110	3	803	308	1221	111	267.7
Total	50	356	35	2444	1438	4238		

Processing Information wpm4238 / 30min.= 141.3
Processing Information wpi 4238 / 50items= 84.8
wpt: words per text *wpi : words per item*

Table 6: Quantity of content input of TEM4 (Time: 15min.)

Section II Part	*Stem No.*	*Stem Words*	*Text No.*	*Text Words*	*Option Words*	*Total Words*	*wpi*	*wpt*
A Conversations	10	87	3	667	102	856	85.6	222
B Passages	10	80	3	566	238	884	88.4	188.7
C News	10	92	5	497	131	720	72	99
Total	30	259	11	1730	471	2460		

Processing Information wpm2460 / 15min.= 164
Processing Information wpi2460 / 30items= 82

should be noticed that: first, "live" human input differs from the "reproduced one", as in a tape recording. Actually, much of the input is not authentically live but reproduced. The news in TEM4, however, may be seen as semi-live input. Second, the selected response here refers to the multiple-choice item while limited response refers to the Dictation in TEM4.

It could be hypothesized that the task types with a "+" mark in the right column require a higher level of difficulty than those at the left column with a "–" mark. As far as the other three task types with a middle indicator are concerned, " / " mark is used.

Four types were, however, proposed by Yang (1998) for classifying these texts: Humanities (H), science and technology (ST) and biomedicine (B). It can be noticed that the type of social sciences (SS) should be added, for there has appeared to be a tendency that the topical type does not confine to the scope of traditional texts for English majors in China, and such texts as economics and politics are not rare in the TEM4 test papers. It is therefore necessary to make this classification.

Table 7 is a summary of the analyses made in this study. For the convenience of discussion, T4 stands for TEM4 and T for TOEFL; thus T4>T suggests that TEM4 has a greater complexity than TOEFL, and T>T4 is the other way round.

The major findings listed above prove that TOEFL listening sub-test is more demanding than its counterpart in terms of test content input (T: 6 > T4: 5).

DISCUSSION

From the analysis of test results, questionnaire responses and test content, it can be seen that TOEFL listening sub-test requires a higher level of communicative language ability in terms of vocabulary and syntax, and the abilities to infer, imagine, analyze, synthesize and judge. However,the TEM4 listening sub-test is more reliable than its counterparts. It has been proved that the item type of one-turn-taking conversations which does not accord with authentic communication has been replaced by more-turn-taking conversations. Findings also show that both groups of subjects are dependent upon question preview, which they believe causes troubles taking TOEFL listening sub-test and does not accord with authentic communication.

Task Characteristics

According to Brindley and Slatyer (2002), who conducted an exploratory study which focused on the effects of task characteristics and task conditions on learners' performance, speech rate and item format exert an impact on task and item difficulty. This justifies the findings in this paper.

Researchers' opinions vary greatly concerning question preview in second language listening tests. Examining tests like TEM, and CET, another test battery administered in China with CET4 aiming at non-English-major students at the relatively lower level and CET6 higher level, and international tests such as TOEFL or IELTS, it is found that there is a great discordance on the issue of questions preview. Cohen (1984) argues that question preview may affect comprehension positively by focusing the student's attention or supplying information about the text, while Marslen-Wilson and Tyler (1980) assume a negative attitude, thinking that previewing questions interferes with subjective comprehension process, and increases the burden on the test takers' attention.

Table 7: Summary of content analysis of both subtests

Component of the analysis	*Detailed items*	*TEM4*	*TOEFL*	*Greater complexity*
Constitution	Interval	5 seconds	12 seconds	T4>T
	Response type	stems printed	stems read	T>T4
	Speech rate	120 wpm	170 wpm	T>T4
	Min./item	0.5	0.6	T4>T
Quantity	Processing	164 wpm	141.3 wpm	T4>T
Task type	% of task +	37.5	33.3	T4>T
	% of task –	54.2	66.7	T4>T
Sentence type	% of simple	67.6	27.5	T>T4
	% of complex	25.4	62.5	T>T4
Topical type	% of H & SS	100	60	T>T4
	% of ST & B	0	40	T>T4
				T4: 5 T: 6

Findings from this paper contrast with those from other researchers. Buck (1991b) holds that it may have no significant effect and Wagner (2013) believes that access to test questions does not affect test-taker performance.

However, findings from the paper show that both groups of subjects are dependent upon question preview, which they believe causes troubles taking TOEFL listening sub-test and does not accord with authentic communication. Another research conducted by Wang (2011) to examine the question preview of the TEM-4 listening subtest indicates again that question preview weakens the validity, and exerts a negative impact on the teaching of listening and on the development of students' listening competence.

The Assessment Implications

The paper has some important implications for the second language listening assessment. The first one concerns the construct validity of listening tests. As Brunfaut and Revesz (2015) claim, second language listening task difficulty correlates significantly with indicators of phonological, discourse, and lexical complexity and with referential cohesion.If a given test is to assess the test takers' ability to comprehend the target language in real-world context, then texts should be included that display authentic features of natural spoken language in terms of phonological, lexical and grammatical characteristics.

The second implication is the consequential aspect of the construct validity (Wagner 2014). An assessment has its positive or negative impact on stakeholders, known as washback effect. As Messick (1989, 1996) argues, social and educational impact of an assessment should be taken into account when tests are designed and developed. If large-scale high-stakes assessments enable test takers to preview test questions (Wang 2011) or use "scripted texts" (Wagner 2014), instructors are unwilling to teach how to process this kind of texts in classroom and second language learners get less motivated to be exposed to these texts beyond classroom.

The Pedagogical Implications

Thus, the pedagogical implication is that instructors should provide more authentic and natural spoken texts or unscripted texts for second language learners instead of scripted or artificial language. As Wagner (2014) argues, by including unscripted texts on these assessments, one can make more valid inferences about test takers' listening ability in that real-world communicative domain.

CONCLUSION

Data comparisons indicate that TOEFL listening sub-test is more demanding than its counterpart and requires a higher level of communicative language ability. TEM4 listening sub-test is, however, more reliable than that of TOEFL. In addition, the test content analysis and its findings prove the above conclusions.

It is also suggested that stems in TEM4 listening sub-test printed on the test book should be removed so as to make test takers unable to preview questions, which will bring about beneficial backwash effects on teaching of listening comprehension and be conducive to improving students' listening proficiency.

Therefore, findings indicate that the research mode, based on the framework of CLA and TMF proposed by Bachman, proves its reliability and validity, which is conducive to furthering language testing studies.

RECOMMENDATIONS

Modern language tests are designed to measure test taker's communicative language ability in real-world settings with more integrated tasks. The proposed research mode is believed to work well if well conceived integrated tasks like reading-to-write or video-and-audio-mediated are designed and if raters factors and test takers' perceptual awareness are to be taken into account.

LIMITATIONS OF THE STUDY

Limitations on the present study are unavoidable. So far as the experimental design is concerned, the limitation is that the first experiment might exert an influence on the second one as two experiments are conducted one after another. Under this circumstance, the subjects may, on the one hand, perform better in the second test than in the first due to practices from the first test or adaptation to aural input; on the other hand, they may also perform worse than in the first test due to the absent-mindedness caused by fatigue. Secondly, the data from the TEM4 test results are likely to be relatively high because of the subjects' test familiarity, which may influence the reliability of the experiment.

REFERENCES

Alderson JC, Clapham C, Wall D 2000. *Language Test Construction and Evaluation*. Beijing: Foreign Language Teaching and Research Press.

Bachman LF 1990. *Fundamental Considerations in Language Testing.* Shanghai: Shanghai Foreign Language Education Press.

Bachman LF, Palmer AS 1997. *Language Testing in Practice.* Shanghai: Shanghai Foreign Language Education Press, pp.124-125.

Brindley G, Slatyer H 2002. Exploring task difficulty in ESL listening assessment. *Language Testing*, 19(4): 369-394.

Brunfaut T, Revesz A 2015.The role of task and listener characteristics in second language listening. *TESOL Quarterly,* 49(1): 141-168.

Buck G 1991. The testing of listening comprehension: An introspective study. *Language Testing*, 8(1): 67-91.

Cohen AD 1984. On taking language tests: What the students report. *Language Testing*, 1(1): 70-81.

Harding L, Alderson JC, Brunfaut T 2015. Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing*, 32(3): 317-336.

Marslen-Wilson W, Tyler LK 1980. The temporal structure of spoken language understanding. *Cognition*, 8(1): 1-71.

Messick S 1989. *Validity.* In: R Linn (Ed.): *Educational measurement*. New York: American Council on Education and Macmilla, pp. 13-103.

Messick S 1996. Validity and washback in language testing. *Language Testing*, 13(1): 242-256.

Richards JC 2002. *Longman Dictionary of Language Teaching and Applied Linguistics*. Beijing: Foreign Language Teaching and Research Press.

Rubin J 1994. A review of second language listening comprehension research. *Modern Language Journal*, 78(2): 199-221.

Spolsky B 1995. *Measured Words.* Shanghai: Shanghai Foreign Language Education Press.

Wang W 2011. Research on question preview of the TEM-4 listening sub-test. *Journal of Xi'an Polytechnic University*, 25(3): 440-442.

Wagner E 2013. An investigation of how the channel of input and access to test questions affect L2 listening test performance. *Language Assessment Quarterly,* 10(2): 178-195.

Wagner E, Toth PD 2014. Teaching and testing L2 Spanish listening using scripted vs. unscripted texts. *Foreign Language Annals*, 47(3): 404-422.

Wood R 2001. *Assessment and Testing: A Survey of Research.* Beijing: Foreign Language Teaching and Research Press.

Wray A, Trott K, Bloomer A, Reay S, Butler C 2001. *Projects in Linguistics: A Practical Guide to Researching Language*. Beijing: Foreign Language Teaching and Research Press.

Yang HZ, Weir C 1998. *A Validation Study of the College English Test CET 4 and CET 6.* Shanghai: Shanghai Foreign Language Education Press.

Zhou YM 2004. *A Comparability Study of CET-6 and TEM-4*. Shanghai: Shanghai Foreign Language Education Press.