ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 7, No. 4, pp. 690-699, July 2016
DOI: http://dx.doi.org/10.17507/jltr.0704.08

Investigating the Construct Validity of Communicative Proficiency in TEP (Oral) at Level B

Weijie Zhou
Department of English Education, Beijing International Studies University, China

***Abstracts*—TEP (oral), a shortened name of Test of English Proficiency on Speaking, is a three-level English oral test that aims to assess the oral English proficiency of university students in those universities administered and supervised by Beijing Municipal Commission of Education. The development of TEP (Oral) has undergone several years and is still being improved. Based on the test scores from senior students of non-English majors in Beijing International Studies University, the study in this paper attempts to investigate the construct validity of the test on students' communicative proficiency. It is hoped that the investigation can shed insights on the improvement of the test and provide empirical insights for language teaching in the future. In this study, the investigation is done using both Classical Test Theory and Multi-facet Rasch Model.**

***Index Terms*—speaking tests, language proficiency, construct validity, Classical Test Theory, Multi-facet Rasch Model**

I. Introduction

The language teaching in China in the past several decades has witnessed the drastic shift from previously the Grammar Translation Method only to the current presence of different teaching approaches and methods where, instead of the focus on reading and linguistic structure, the teaching and learning in listening, speaking, and writing have been given adequate attention. In addition, more teaching activities have been conducted in language teaching and learning to cover sociolinguistic and social cultural elements.

The development of tests to promote educational reforms has gained increasing popularity in the past decades (James, 2000; Chapman & Sydney, 2000) nationwide and worldwide. Moreover, McNamara (1996) observed that since the early 1990s, in language assessment more traditional paper-and pencil tests involving multiple choice questions has been supplemented or even replaced with performance assessments where language learners have to demonstrate practical command of skills acquired. However, studies have shown that performance assessments can also bring about some unexpected effects on test scores due to some factors in performance assessments, such as raters and tasks. (Bachman & Palmer, 1996; Brown, Hudson, Norris, & Bonk, 2002). Therefore, a test, a big-scale test in particular, should undergo a long process of development.

TEP (oral) is a shortened name of Test of English Proficiency on Speaking. It is a three-level English oral test that aims to assess the oral English proficiency of university students in those universities administered and supervised by Beijing Municipal Commission of Education. A performance assessment to understand to what degree the students' communicative proficiency has met the requirements and criteria of university English teaching and learning issued by Chinese National Ministry of Education, TEP (Oral) is expected, by having a closer look at the university students' proficiency in speaking, to help English teachers with the understanding of students' language proficiency and provide them with empirical guidance in making practical changes in language teaching.

TEP (Oral) sets three levels as Level A, B and C, with Level A as the highest level of proficiency and Level C the lowest. The tests are paired in that there are two examiners (one as the interlocutor and the other as assessor) and two test-takers to accomplish interactive tasks. Different from TOEFL and IELTS, the test items in each level, except for the first part, warming-up questions that are not included in the scoring, consist of two speaking tasks and are designed in varied forms. Though the items in the scoring scales are the same, the items are given different weightings at different levels. In this paper, the study focuses on TEP (Oral) at Level B. The purpose of this paper is to investigate the Construct Validity of communicative competence in TEP (Oral) at Level B through quantitative study.

Construct validity refers to the degree to which a test measures what it claims to be measuring, specifically whether a test measures the intended construct. (Brown, J. D. 1996; Cronbach, L. J. & Meehl, P.E., 1955) To understand the construct validity of TEP (Oral) at Level B, 5 research questions are included. 1) To what extent has the internal consistency reliability of the test reached? 2) To what extent can the inter-rater consistency reliability of the test achieve? 3) To what extent can the scores reflect construct validity? 4) How has the weighting of different items of communicative proficiency been appropriately set? 5) What is the relative contribution of multiple sources of variation (e.g. test-takers' language ability, task difficulty, and the raters' rating scores) to the total score variability in TEP (Oral)? To answer these questions, the study in this paper has used SPSS 22.0 and MINISTEPS 3.74.0.

There are three main parts in this paper: understanding the nature of Communicative Proficiency; the test items and scoring scales of TEP (Oral) at Level B; and the quantitative study of the construct validity of communicative proficiency in TEP (Oral) at Level B.

II. UNDERSTANDING THE NATURE OF COMMUNICATIVE PROFICIENCY

According to Yalden (1997), in traditional approaches to language teaching, proficiency of a language learner is viewed as unitary and described in the degree of the language learners' mastery of "structures" – that is, of the phonology, morphosyntax, and lexicon of the target language.

In 1960s Chomsky's (1965) distinction between linguistic competence and linguistic performance has brought about a revolution in understanding language and language learning. It has stimulated new developments in linguistic study and classroom language learning. However, it was generally believed in 1970s that linguistic performance proposed from Chomsky had only psychological constraints on performance and Chomsky ignored all aspects of social interaction (Hymes, 1972). Therefore, Hymes (1972), for example, claimed that a different theory of language was needed by individuals involved in language development. In such a theory, competence would be called 'communicative competence' because of the inclusion of interactional competence. Hymes' theory (1972) of communicative competence has linked linguistic theory to a more general theory of communication and culture, and he involved judgments of four kinds: possibility, feasibility, appropriateness and actual performance. Hymes' theory has thus suggested that grammaticality is only one of four sectors of communicative competence, far different from Chomsky's belief that grammaticality was competence.

Halliday and Hasan (1976, 1989) and Halliday (1994), from a socially-oriented perspective of language, have advocated the notion that language is communication-based, and not primarily form-based. According to Halliday (1994), language is made up of a small set of universal communicative functions that includes experiential (relating to experiences), interpersonal (relating to social relationships) and textual (relating to structure) functions.

Hymes' and Halliday's work on communicative competence have had a great impact on the formation of communicative proficiency in that models with communicative language ability(CLA) have emerged.

Canale and Swain (1980), and later Canale(1983), put forth a model that lists a total of four areas of knowledge and skill in communicative competence, including "grammatical competence (mastery of the language code); sociolinguistic competence (appropriateness of utterances with respect both to meaning and form); discourse competence (mastery of how to combine grammatical forms and meanings to achieve unity of a spoken or written text); and strategic competence (mastery of verbal and non-verbal communication strategies used to compensate for breakdowns in communication, and to make communication more effective)" (Canale, 1983, p. 9-10).

Based on Canale and Swain's (1980) and Canale's(1983) scheme of linguistic, sociolinguistic, and strategic components, Lyle Bachman (1990) introduced a model of communicative language ability (CLA) that includes three main components: language competence, strategic competence and psychophysiological mechanism. Language competence is further divided into organizational competence and pragmatic competence, and strategic competence is defined in terms of metacognitive strategies, such as goal setting, assessment and planning. Bachman's Model of CLA has been very influential in language learning and testing. Taking Bachman's Model as the theoretical basis for the construction of its speaking tests, TEP (Oral) attempts to measure students' communicative proficiency in terms of their grammatical competence, pragmatic competence and strategic competence.

III. THE TEST ITEMS AND SCORING SCALES OF TEP (ORAL) AT LEVEL B

The goal of TEP (Oral) is set to see whether the learners' proficiency has met the requirements at 'the relatively higher level' issued by documents from the Commission of English Teaching in Higher Education, the Ministry of Education(2012). In other words, for test-takers at TEP (Oral) Level B, they should demonstrate their competence in accomplishing with adequate fluency on common topics (e.g. campus life, environmental protection, etc.), stating factual information, describing events, reasoning and expressing their personal views.

TEP (Oral) at Level B comprises of three parts that are expected to be finished in around 12 minutes. In the first part, which lasts for 1 minute, each test-taker is given one question so as to gain some familiarity with the interlocutor's voice, volume, pace, pitch, pronunciation and intonation, etc. In return, the assessor and the interlocutor get the first impression of the two test-takers' performance from their answers to the questions.

The second part, retelling of a passage, is segmented into three subparts, with each part about 2 minutes. First the interlocutor hands out each test-taker a piece of paper with a different passage of about 150 words. Each test-taker is required to read his/her passage and take some notes on the paper prepared on the desk within 2 minutes. In the second subpart, one test-taker starts his/her retelling of the passage with the only help from the written notes, and the retelling is finished in 1 minute. After the retelling, the other test-taker, based on what he/she has listened to his partner, asks a related question and waits for the answer. The same procedure is repeated by the next test-taker in the third subpart.

In the final part of TEP (Oral) at Level B, the two test-takers work together on a task on reciprocal basis. Each test-taker prepares on his/her own for two minutes with a clip of paper on which the same detailed instructions are given. When the required time is up, the two test-takers are assessed with their 3-minute co-performance through oral discussion.

The test items are designed to assess the learners' ability in achieving life-related tasks with English in speaking context. The questions posed at the first part are related to test-takers personal life experiences. The tasks in the second and third part are either real-world related or pedagogically related, covering such areas as education, employment, law, trade and economy, sports, travelling, food and health, famous people, natural environment and habitat, and Internet-related experiences.

In TEP (Oral) at Level B, each test-taker's score comes from the holistic evaluation from the interlocutor (accounting for 40% of the total score) and the analytic evaluation from the assessor (accounting for the remaining 60% of the total score). The analytic evaluation, in reference to Bachman model of CLA (1990), is described in four items, representing major dimensions of communicative proficiency: Communicative Effect, Content and Organization, Pronunciation and Intonation, Syntax and Vocabulary. Each item is rated on a 5-point-scale, ranging from 1 as the lowest to 5 as the highest, with descriptors at each point of the scale.

Communicative Effect in the analytic evaluation of TEP (Oral) at Level B is considered relevant to the perspective of strategic competence from Bachman (1990). A test-taker's communicative effect is measured in terms of fluency, appropriateness, interactivity and the use of communicative strategies, such as physiological mechanisms. In Pronunciation and Intonation, a test-taker is assessed at the criterion of whether his/her speaking is comprehensible or not. In Syntax and Vocabulary, the range of 1 to 5 is given based on the variety of sentential structures and appropriate use of words.

Unlike Bachman's CLA model, the analytic evaluation of TEP (Oral) has used Content and Organization as the combination of textual competence from grammatical competence and illocutionary and sociolinguistic competence from pragmatic competence. A successful demonstration of Content and Organization in retelling of a passage is seen from not only the coherence of the retelling, but also the range of the content a test-taker is required to cover, besides the use of cohesion. In the third part of TEP oral test at Level B, the test takers' performance in the topic discussion is evaluated in terms of the degree of appropriate conveyance and interpretation of implied sociolinguistic, sociocultural, and psychological meanings encoded in high-context language use.

TEP (Oral) at Level B, studied in this paper, was conducted on the weekend of Nov. 1st, 2014. Even though TEP (Oral) at Level B has been administered several times before, a training session has still been required as prerequisites for 36 raters, who were English teachers coming from Department of English Education in Beijing International Studies University. The session was to make sure that the raters were aware of the procedure and instructions involved in the test and would follow the scoring criteria. The senior undergraduate students were encouraged to take the test voluntarily. On the day of the speaking test, 36 trained raters were divided into 18 pairs. There were 254 senior undergraduate students who volunteered to take part in the test at Level B. In the test, every 4 pairs of test-takers in each testing room was given a set of test package, including the instructions and the timer for the raters; two sheets of blank paper, two pens and the instructions with the speaking tasks for the test-takers. After 4 pairs were done, a new set of test package with different tasks was given to the raters.

IV. THE QUANTITATIVE STUDY OF CONSTRUCT VALIDITY OF COMMUNICATIVE PROFICIENCY IN TEP (ORAL) AT LEVEL B

The study of construct validity in this paper has attempted to examine whether the test has reflected what the test designers have intended to achieve through the test. The questions are: 1) To what extent has the internal consistency reliability of the test reached? 2) To what extent can the inter-rater consistency reliability of the test achieve? 3) To what extent can the scores reflect construct validity? 4) How has the weighting of different items of communicative proficiency been appropriately set? 5) What is the relative contribution of multiple sources of variation (e.g. test-takers' language ability, task difficulty, and the raters' rating scores) to the total score variability in TEP (Oral)?

According to Classical Test Theory (CTT), the internal consistency reliability of the test is calculated through Cronbach's alpha. From Table1, the high internal consistency reliability of TEP (Oral) at Level B can be seen from high corrected Item-Total Correlation and Cronbach's Alpha if Item Deleted, which can thus prove the reliability of the test.

TABLE 1
ITEM-TOTAL STATISTICS

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
communicative effect	17.4406	5.857	.870	.93
text retelling	17.5075	6.275	.802	.945
topic discussion	17.4878	5.996	.815	.945
pron. & into.	17.4465	6.429	.764	.949
vocab & gram	17.5154	6.221	.860	.939
analytic total	17.4626	6.144	.995	.926

In TEP (Oral) at Level B, each test-taker's total score comes from the holistic evaluation from the interlocutor (accounting for 40% of the total score) and the analytic evaluation from the assessor (accounting for the remaining 60% of the total score). To understand whether the score from holistic evaluation is correlated to the one from the assessor, a Spearman rank-order correlation is used to calculate the inter-rater reliability in each pair. (See Table 2)

The inter-rater reliability for each pair of judges is listed in Table 2 in the order from the highest one to the lowest. In this table, high correlation coefficients can indicate that both raters have followed the criterion simultaneously while those low correlation coefficients, especially the last 5 ones in Table 2, have displayed the vast disagreement between the raters within the pairs. Therefore, 40 percent of the holistic evaluation and 60 percent of the analytic assessment seem to be a remedial treatment to ensure the fairness of a test-taker's score in the test. In fact, the Spearman's rank correlation coefficient between the holistic evaluation from all the interlocutors and the analytic evaluation from all the assessor still remained at 0.629, thus the inter-rater reliability of the test can still be considered with fairly high correlation. Moreover, on the premise that all the raters have taken the training session before the test, the variance of the inter-rating has confirmed the importance of the training sessions to the raters.

TABLE 2
INTER-RATER RELIABILITY AMONG EACH PAIR OF JUDGES

Pair numbers	Correlation coefficient of different pairs (Spearman's rho)
7	0.973
2	0.971
5	0.946
13	0.858
10	0.839
16	0.836
1	0.828
12	0.719
11	0.692
18	0.688
14	0.676
4	0.645
8	0.586
15	0.466
6	0.445
17	0.253
3	0.226
9	0.036

**. Correlation is significant at the 0.01 level (2-tailed).

After calculating the inter-rater reliability and Cronbach's alpha, Factor Analysis is used to examine the construct validity (Qin 2003). To get the construct validity, an exploratory factor analysis is done through SPSS 22. The results are presented in Table 3, 4 and 5 and Graph 1. Through KMO and Bartlett's Test, we can see the significant level is .000, which is lower than 0.05 and therefore, KMO and Bartlett's Test can serve as the evidence to support the existence of common factors and the feasibility of exploratory factor analysis.

TABLE 3
KMO AND BARTLETT'S TEST (P<0.05)

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.886
Bartlett's Test of Sphericity	Approx. Chi-Square	973.681
	df	10
	Sig.	.000

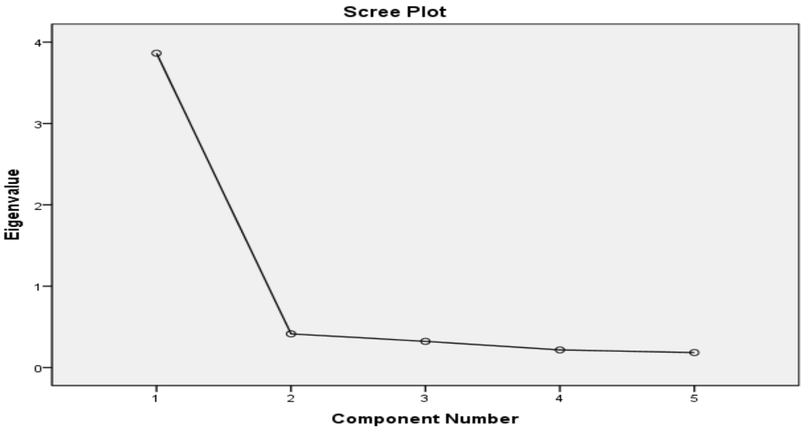
Through communalities in Table 4, we can see high correlations between the five items of the rating scales. From Total Variance Explained in Table 5 and the Scree Plot in Graph 1 it can be seen that the five items are functioning together to contribute to the test-takers' communicative proficiency.

TABLE 4
COMMUNALITIES EXTRACTION METHOD: PRINCIPAL COMPONENT ANALYSIS

	Initial	Extraction
communicative effect	1.000	.837
text retelling	1.000	.750
topic discussion	1.000	.771
vocab & gram	1.000	.822
pron. & into.	1.000	.684

TABLE 5
TOTAL VARIANCE EXPLAINED (EXTRACTION METHOD: PRINCIPAL COMPONENT ANALYSIS)

	Initial Eigenvalues			Extraction Sums of Squared Loadings	
Component	Total	% of Variance	Cumulative %	Total	% of Variance
1	3.864	77.270	77.270	3.864	77.270
2	.414	8.287	85.557		
3	.322	6.433	91.990		
4	.217	4.333	96.323		
5	.184	3.677	100.000		



Graph 1 Scree Plot

To examine to what extent the weighting of different items of the scoring scales has been appropriately set, the regression analysis is done through SPSS 22(Xu Hongchen, 2013).

After the regression linear analysis in stepwise method (Table 5), the communicative proficiency can be written as: communicative proficiency= 0.195communicative effect + 0.150text retelling + 0.139 topic discussion+ 0.302 pronunciation &intonation+ 0. 202 vocabulary & Grammar. Therefore, if we convert the equation of the communicative proficiency in the regression linear analysis to the weighting, the weighting would be counted roughly as: 0.2, 0.15, 0.15, 0.3, 0.2. If we compare the calculated weighting with the constructed weighting of the five different items in scoring scales: 0.2communicative effect; 0.15 text retelling; 0.15 topic discussion; 0.3 pronunciation & intonation; and 0.2 vocabulary &grammar, we can discover a fairly good match between the two.

To answer Question 5 in this study, Multi-facet Rasch Models (MFRM) – MINISTEPS 3.74.0 is used. The reason to use MEFM is that a performance assessment, in reality, can be affected by multiple sources of variance, such as raters and task difficulties. Therefore, CTT fails to take into account of the possibility of the interaction of different sources of error from raters and tasks, and may cause the construction of irrelevant variance. In contrast, MFRM can provide more detailed information about interactions between the interaction of a specific rater with a certain test-taker or a task and improve the precision of the investigation of construct validity. (Lynch & McNamara, 1998; McNamara & Roever, 2006; Grabowski, K. C., 2009; Rasch G. 1960; Fan Jingsong & Ji Peiying, 2015; Wang Jimingy, 2002; Lincare, J. M. 2012)

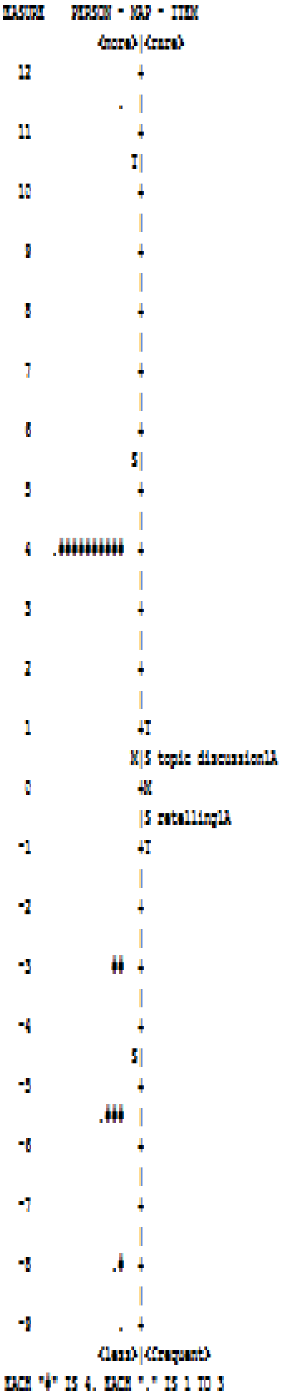
In addition, as the result from the factor analysis in answering the construct validity of TEP (Oral) at Level B has announced that there is only one principal common factor among the 5 items of the rating scale, the result supports MFRM positively in that the result has met the requirement that the 5 items are one dimension in nature and MFRM can be used.

TABLE 8
MULTIPLE LINEAR REGRESS: IMPORTANT STATISTICS (COEFFICIENTS')

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.942	.076		12.347	.000
	communicative effect	.727	.021	.907	34.132	.000
2	(Constant)	.338	.052		6.493	.000
	communicative effect	.471	.017	.587	27.777	.000
	pron. & into.	.428	.019	.469	22.211	.000
3	(Constant)	.215	.036		6.008	.000
	communicative effect	.313	.015	.391	21.519	.000
	pron. & into.	.391	.013	.428	29.649	.000
	topic discussion	.233	.013	.290	17.402	.000
4	(Constant)	.128	.025		5.033	.000
	communicative effect	.244	.011	.304	22.320	.000
	pron. & into.	.340	.010	.372	35.263	.000
	topic discussion	.170	.010	.212	16.939	.000
	vocab & gram	.211	.013	.230	16.452	.000
5	(Constant)	.052	.011		4.575	.000
	communicative effect	.195	.005	.243	38.933	.000
	pron. & into.	.302	.004	.331	69.080	.000
	topic discussion	.139	.004	.173	31.085	.000
	vocab & gram	.202	.006	.221	36.152	.000
	text retelling	.150	.005	.169	32.545	.000

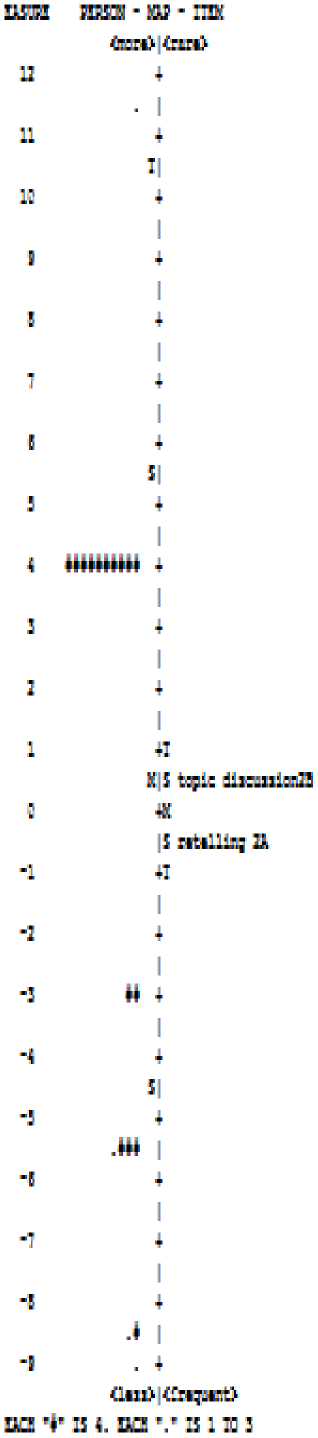
Multiple Linear Regress: Important Statistics (n=254), P<0.05

WINSTEPS 3.74.0 is employed here to understand what relative contribution of multiple sources of variation (e.g. test-takers' language ability, task difficulty, and the raters' rating scores) is to the total score variability in TEP (Oral) at Level B.



```
MEASURE   PERSON - MAP - ITEM
                <more>|<rare>
  12                  +
                    . |
  11                  +
                     T|
  10                  +
                      |
   9                  +
                      |
   8                  +
                      |
   7                  +
                      |
   6                  +
                     S|
   5                  +
                      |
   4      .#########  +
                      |
   3                  +
                      |
   2                  +
                      |
   1                  +T
                     M|S topic discussion1A
   0                  +M
                      |S retelling1A
  -1                  +T
                      |
  -2                  +
                      |
  -3                ##  +
                      |
  -4                  +
                     S|
  -5                  +
                .###  |
  -6                  +
                      |
  -7                  +
                      |
  -8                .#  +
                      |
  -9                 .  +
                <less>|<frequent>
EACH "#" IS 4. EACH "." IS 1 TO 3
```

Chart 1 Set 1



```
MEASURE   PERSON - MAP - ITEM
                <more>|<rare>
  12                  +
                    . |
  11                  +
                     T|
  10                  +
                      |
   9                  +
                      |
   8                  +
                      |
   7                  +
                      |
   6                  +
                     S|
   5                  +
                      |
   4      ##########  +
                      |
   3                  +
                      |
   2                  +
                      |
   1                  +T
                     M|S topic discussion2B
   0                  +M
                      |S retelling 2A
  -1                  +T
                      |
  -2                  +
                      |
  -3                ##  +
                      |
  -4                  +
                     S|
  -5                  +
                .###  |
  -6                  +
                      |
  -7                  +
                      |
  -8                  +
                  .#  |
  -9                 .  +
                <less>|<frequent>
EACH "#" IS 4. EACH "." IS 1 TO 3
```

Chart 2 Set 2

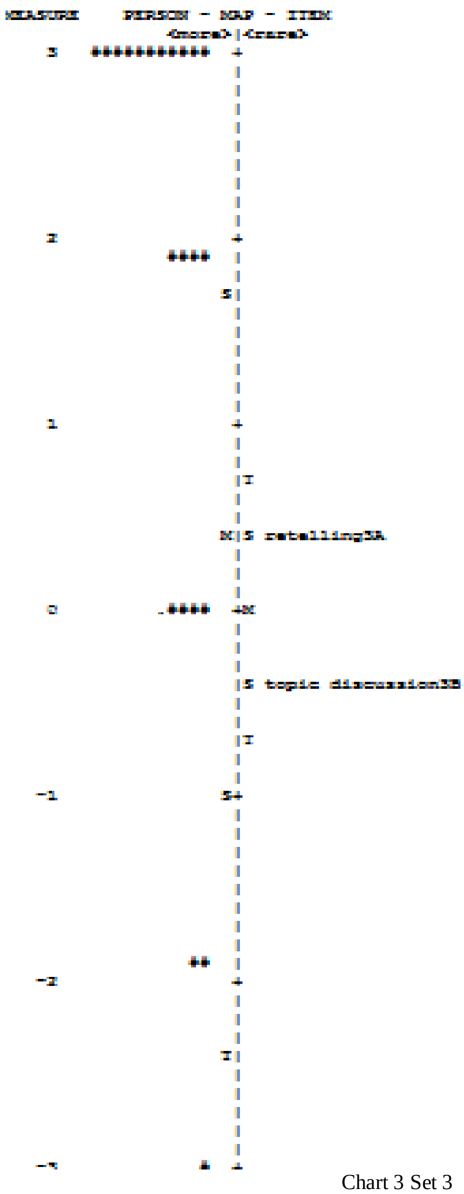


Chart 3 Set 3

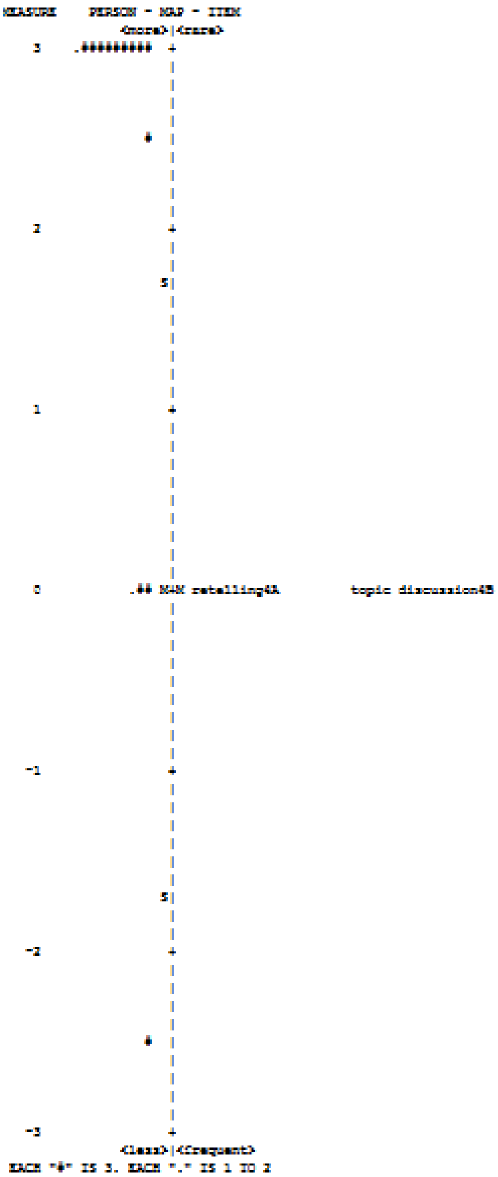


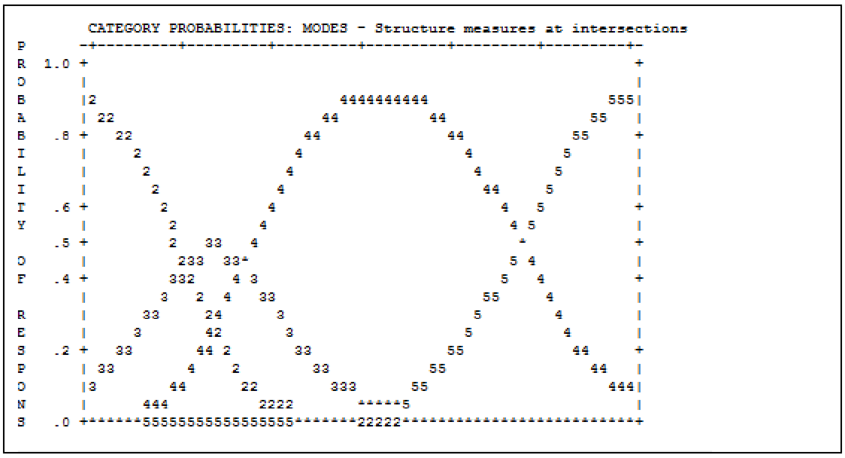
Chart 4 Set 4

TABLE 10

SUMMARY OF CATEGORY STRUCTURE. MODEL="R"

```
|CATEGORY      OBSERVED|OBSVD SAMPLE|INFIT OUTFIT|| ANDRICH |CATEGORY|
|LABEL SCORE COUNT %|AVRGE EXPECT|    MNSQ    MNSQ||THRESHOLD| MEASURE|
|-------------------+------------+------------++---------+--------|
|   2      2         110      7| -1.75 -1.72|      .99       .95||    NONE      |( -4.15)| 2
|   3      3         434    28| -1.30 -1.26|      .96       .83||      -2.88 |     -2.26 | 3
|   4      4         964    63|     -.24     -.28|    1.00       .99||      -1.64 |       1.45 | 4
|   5      5           16      1|     -.27* 1.05|    1.41       .69||        4.52 |(     5.62)| 5
|-------------------+------------+------------++---------+--------|
|MISSING          204     12|     -.84            |                       ||                |              |
--------------------------------------------------------------------
```

TABLE 11
CATEGORY PROBABILITIES: MODES - STRUCTURE MEASURES AT INTERSECTIONS



```
        CATEGORY PROBABILITIES: MODES - Structure measures at intersections
P      -+---------+---------+---------+---------+---------+---------+-
R  1.0 +                                                             +
O      |                                                             |
B      |2                            4444444444                    555|
A      | 22                       44          44                55   |
B   .8 +   22                   44              44            55     +
I      |     2                 4                  4          5       |
L      |      2               4                    4        5        |
I      |       2             4                      44     5         |
T   .6 +        2           4                         4   5          +
Y      |         2         4                           4 5           |
    .5 +         2   33   4                             *            +
O      |          233  33*                             5 4           |
F   .4 +         332    4 3                           5   4          +
       |        3   2  4   33                       55     4         |
R      |      33     24      3                     5        4        |
E      |     3       42       3                   5          4       |
S   .2 +   33       44 2       33               55            44     +
P      | 33        4    2        33           55                44   |
O      |3        44      22        333      55                    444|
N      |      444          2222       *****5                         |
S   .0 +******55555555555555555*******22222**************************+
```

Four charts, which are Chart 1,2, 3and 4, have displayed the distributions of test-takers' abilities and the task difficulty. In each chart, there are three main parts. First, the measures, calculated in logits, measures like a meter rule the test-takers' abilities and the task difficulty. Next to the measures, on the left-handed column list the abilities of test takers, and 'less' refers to the lower abilities of the test-takers, while 'more' the higher abilities; meanwhile, the right-handed column on the chart goes from 'frequent' to 'rare', indicating the increasing difficulty from the bottom to the top.

In Chart 1 and 2, the test-takers' abilities are set from -9 to +5 logits, showing a vast disparity of test-takers' language abilities; meanwhile, the measures of task difficulty have shown that text retelling is more difficult than topic discussion, yet both tasks, text retelling and topic discussion, are rather easy for most test-takers. In Chart 3 and 4, the test-takers' language abilities range from -3 to +3 logits, showing a fairly narrow convergence than the previous 2 charts. Still it can be seen from the last two charts that tasks are easy to the test takers. As for the task difficulty, text retelling in Chart 3 is even more difficult than topic discussion, with a range of nearly 1 logit; however, in Chart 4, text retelling is as difficult as topic discussion. The results from the four charts have thus shown that the tasks in four sets of test package are not quite equivalent in difficulty, and thus calling for more considerations and improvement in constructing the speaking tasks so as to obtain a true picture of test takers' language abilities.

Table 10, together with Table 11, has presented the frequency of score scales the raters have given to the test-takers. The results have shown that among the five scales in each category, four scales are used, and the most frequent ones is 4. The raters' rating scores are relevant to Chart 1,2,3 and 4 in that the test takers outperformed the intended consequences of the speaking tasks. However, it can be inferred from Table 10 and 11 that the raters have used the rating scales appropriately and the scales can differentiate the test-takers in terms of their language proficiency.

V. CONCLUSION

TEP (Oral) is constructed to assess the university students' proficiency in speaking so as to inform English teachers of student performance achievements in language learning and to provide empirical guidance to teachers in making changes in the classrooms.

Using SPSS 22 and MINISTEPS 3.74.0, the study in this paper has examined the construct validity of TEP (Oral) at Level B, which means what the test designers intended to examine has been achieved. The high internal consistency of reliability, the high inter-rater consistency, as well as the results from Factor analysis, have proved the construct validity of TEP (Oral) at Level B in that the five categories and 1-5 points in the rating scales are homogeneous in contributing to the assessment of communicative proficiency. However, the study through MFRM has also indicated that improvements should be done to TEP (Oral) at Level B. As the tasks at Level B are fairly easy for most test-takers' abilities, the tasks of text retellings and topic discussions should be reconstructed so that the tasks can truly reflect the test takers' language abilities. Also, if a qualitative analysis on the test takers' performance is done, the understanding of the test takers' language proficiency can help the test designers with working out more effective speaking tasks.

Furthermore, TEP(Oral) is a three-level speaking test. Therefore, the study of construct validity has even more to do. To gain a fuller picture of the construct validity of TEP (Oral) at three levels, the author of the paper will continue the further investigation of the construct validity of TEP (Oral) at Level A and C quantitatively and qualitatively.

ACKNOWLEDGEMENT

This study is the partial fulfillment of the Project (Project No. – 090031) funded by Social Science Research Program from Beijing International Studies University.

REFERENCES

[1] Bachman, L. (1990). Fundamental considerations in language testing. Oxford, UK:

[2] Bachman, L. and Palmer, A. (1996). Language testing in practice. Oxford, UK: Oxford University Press.

[3] Brown, J. D. (1996). Testing in language programs. Upper Saddle River, NJ: Prentice Hall Regents.

[4] Brown, J. D., Hudson, T., Norris, J., and Bonk, W. (2002). An Investigation of Second language task-based performance Assessments (Technical Report #24). Honolulu, HI: University of Hawaii, Second Language Teaching & Curriculum Center.

[5] Canale, M. (1983). On some dimensions of language proficiency. In J. Oiler (Ed.), *Issues in language testing* (pp.333-42). Rowley, MA: Newbury House.

[6] Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, I,* 1-47.

[7] Chapman, D. W., & Snyder, C. W. (2000). Can high-stakes national testing improve instruction: Reexamining conventional wisdom. *International Journal of Educational Development*, 20, 457–474.

[8] Chomsky, N. (1965). Aspects of the theory of syntax. Cambridge, MA: MIT Press.

[9] Cronbach, L. J.; Meehl, P.E. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin*, 52 (4), 281–302. doi:10.1037/h0040957. PMID 13245896.

[10] Fan Jingsong; Ji Peiying. (2015). Construct validation of an analytic rating scale for speaking assessment. *Foreign Language Education in China.* 8(3), 85-94.

[11] Grabowski, K. C. (2009). Investigating the construct validity of a test designed to measure grammatical and pragmatic knowledge in the context of speaking. http:// www.docin.com (Accessed 29/9/2015).

[12] Halliday, M. A. K. (1994). An introduction to functional grammar (2nd ed.). London: Edward Arnold.

[13] Halliday, M. A. K., & Hasan, R. (1976). Cohesion in English. London: Longman.

[14] Halliday, M. A. K., & Hasan, R. (1989). Language, context, and text: Aspects of language in a socio-semiotic perspective. Oxford, UK: Oxford University Press.

[15] Hymes, D. (1972). On communicative competence. In J. Gumperz & D. Hymes (Eds.) *Directions in sociolinguistics* (pp. 35-71). New York: Holt, Reinhart, & Winston.

[16] James, M. (2000). Measured lives: The rise of assessment as the engine of change in English schools. *Curriculum Journal* 11, 343–364.

[17] Lincare, J. M. (2012). Rasch-Model Computer Programs - Program Manual 3.74.0. http:// www. Winsteps.com (accessed 10/12/2015).

[18] Lynch, B., & McNamara, T. (1998). Using G-theory and Many-Facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15,* 158-180.

[19] McNamara, T. (1996). Measuring second language performance: A new era in language testing. New York: Longman.

[20] McNamara, T., & Roever, C. (2006). Language testing: The social dimension. Maiden, MA: Blackwell.

[21] Qin Xiaoqing. (2003) Quantitative Statistic Analysis in Foreign Language Teaching Research. Wuhan, China: Central China Science and Technology University Press.

[22] Rasch, G. (1960). Probabilistic Models for some intelligence and attainment test. Copenhagen: Danish Institute for Educational Research.

[23] University English Teaching Guidelines. http://www.doc88.com/p-3107130241960.html (accessed 12/8/2015).

[24] Wang Jimingy. (2002). A study of the scoring of three types of oral test items. http:// www.cnki.com. DOI: 10. 13724/j.cnki.ctiw. 2002.04.009 (accessed 12/9/2015).

[25] Xu Hongchen. (2013). Learning Statistics from Examples of Second Language Research. Beijing: Foreign Language Teaching and Research Press.

[26] Yalden, J. (1997). Principles of Course Design for Language Teaching. New York: Elsevier.

Weijie Zhou is currently Associate Professor in Beijing International Studies University. She earned a Master's degree in TESOL in Beijing International Studies University and a Master's degree in Applied Linguistics in Southern Queensland University, Australia.

She has co-authored 6 books and textbooks in Translation and Language teaching. Also she has published more than 10 research papers on language teaching and translation. Her main research interests are Applied Translation and Applied linguistics.