ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 9, No. 5, pp. 939-945, September 2018
DOI: http://dx.doi.org/10.17507/jltr.0905.06

The SPSS-based Analysis of Reading Comprehension—Take Grade Eight English Mid-term Test for Example

Yingying Jin
University of Shanghai for Science and Technology, China

Xiaowen Qi
University of Shanghai for Science and Technology, China

***Abstract*—Based on some language testing theories, an analysis of an English Mid-term Examination of grade eight students of JingZhi Middle School is made in this paper. By means of SPSS statistical software, the former study firstly makes a whole analysis of the test paper, which covers descriptive statistics, reliability and validity. Subsequently, on the basis of the former study, this study mainly makes an analysis of the relationship of all the items in reading comprehension from the perspective of facility value, discrimination index, reliability and validity. The research aims to find some problems in reading comprehension in this test paper. Thus, according to the results of this analysis, the quality of test papers can be improved and some advice can be given to language teaching.**

***Index Terms*—reliability, validity, F.V., D.I., SPSS**

I. Introduction

With the development of society and technology, language testing becomes more and more important and popular in different kinds of fields,especially for language teaching. The Test paper is a necessary way for examination. Thus, the analysis of test papers can be regarded as a significant part of language testing, which will provide a guide for teachers and students (Cheng, 2015). An accurate analysis of test papers can improve teachers' teaching efficiency and students' learning ability. Besides, it will instruct teachers to make high-quality test papers to test students' ability.

II. Basic Items in Language Testing

Facility value (F.V.) refers to different levels of the difficulty of test papers. The range of facility value is between 0 and 1 and the ideal range of facility value is between 0.3 and 0.7. Discrimination index (D.I.) refers to different students in different degree, which is very important in maintaining test reliability. The range of D.I. is between -1 and 1, and its ideal range is above 0.35. F.V. and D.I. are only used for objective items.

Reliability refers to the consistency of scores. (Bachman & Palmer, 1996) There are mainly4 methods to examine reliability of scores, including test-retest method, parallel-form method, split-half method and coefficient α method. Among all of these methods, coefficient α is the most commonly used one. And this paper chooses coefficient α to test the reliability of this English test paper. Validity refers to the appropriateness of a test and it is considered as a measure of assessment (Bachman, 1999). Many measures are used to examine the validity of a test, which covers face validity, content validity, concurrent validity, predicative validity, response validity and construct validity. This paper is mainly focused on construct validity, which refers to language ability or competence under question. It is something abstract, including proficiency and strategy skills. Besides, it is related to score interpretation. In general, validity and reliability are equally necessary in language testing. However, it is difficult to make a balance between reliability and validity. In fact, validity is firstly ensured and then reliability. Reliability is also considered as an aspect of validity (Bachman & Palmer, 1996).

SPSS (statistical package for the social sciences), as a statistic tool, is widely used to analyze the test scores in language testing. It is easy to handle and the result is reliable to analyze a test paper (Zou & Dai, 2012).

III. The Process of the Research

A. The Question of the Research

The research has been divided into two parts. One is the whole analysis of the test paper, which has been finished and published. And the other one is the analysis of Reading Comprehension, which is closely related to other sections based on the results of the former study. This paper firstly tries to make a summary on the basis of the former study , which includes descriptive statistics, reliability and validity of this English mid-term test paper with the use of SPSS software.

© 2018 ACADEMY PUBLICATION

Then, a further detailed study of Reading Comprehension will be made to focus on the following aspects in this study: F.V., D.I., reliability and validity. Because the first section of the analysis of the whole paper has been published, so this paper mainly is focused on the analysis of Reading Comprehension. The purpose of this research is to find some problems in this test paper, developing a highly qualitative test paper and improving teachers' teaching and students' learning.

B. The Research Object and Sample of the Research

The object of the study are 30 students from Grade 8 Class 1 in JingZhi Middle School. The sample of the research is students' English mid-term test paper. The valid samples are 30. There are 9 parts in this test paper: Listening Comprehension, Multiple Choice Question, Cloze, Reading Comprehension, Translation between English and Chinese, Filling in the blank, Rewriting the sentences, Translation from Chinese to English and Writing. As shown in the table 1, the types of items are diversified. And the amounts of the subject items are more than the object items.

TABLE 1.
THE TYPES OF ITEMS

	Listening Comprehension	Multiple Choice Question	Cloze	Reading Comprehension	Phrases Translation	Filling in the blank	Rewriting the sentences	Sentences Translation	Writing	Sum
Scores	25	10	10	30	5	5	10	10	15	120
Numbers of Items	25	10	10	15	10	10	5	5	1	91

C. The Methods and Processes of the Research

The research uses SPSS to collect the data and then makes a detailed analysis of Reading Comprehension. Firstly, make a summary of the former study on the whole analysis of the test paper. Then, based on the former study, make a concrete analysis of the Reading Comprehension, calculate the F.V. and D.I. of all the items of Reading Comprehension and collect the data of reliability and validity of Reading Comprehension in this study. Because the first section of the analysis of the whole paper has been finished, so this paper mainly pays attention to the analysis of Reading Comprehension.

IV. RESULTS AND DISCUSSION

A. The Summary of the Analysis of the Whole Test Paper in Former Study

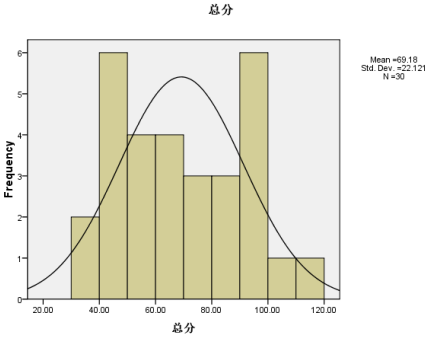
a. The analysis of descriptive statistics

Table 2 reflects the descriptive statistics of the whole test paper. From the following table, in this class, the mean of test is 69.18, which demonstrates that most students can pass the exam. But the level of the whole class is not too high. Maybe not all the students can master the knowledge what they have learned in the past term. The median is 63.75 and the mode is 45.5. The maximum is 112 while the minimum is 38.5, so the range is 73.5. It indicates that the scores of different students are of largely variance. The individual difference of the whole class is a big problem. The S.D. is 22.1, which is too big for the whole class. The higher the S.D. is, the bigger the gap is. It also demonstrates that students in this class have an extremely large difference in their English level. Considering the whole class, the scores of the students are relatively unstable. The polarization between high and low scores students is serious.

TABLE 2.
DESCRIPTIVE STATISTICS

N (Valid/Missing)	Mean	Median	Mode	Std. Deviation	Range
30/0	69.18	63.75	45.50	22.12	73.50

TABLE 3.
HISTOGRAM OF SCORES



© 2018 ACADEMY PUBLICATION

From the above histogram, the scores of the whole class correspond to the normal distribution. By calculation, the skewness is 0.29 and the value of kurtosis is -1.22. They are almost between -2 and 2, which also indicates a reasonable normal distribution. Normal distribution is prerequisite in language testing, for it is the basis for a further analysis. Thus, the items of the test paper are relatively reasonable to test students' ability of the past half-term knowledge. Besides, the most two frequent scores are mainly centralized on 40-50 and 90-100. About 12 students (nearly half of the class) cannot pass the exam, whose scores are below 60. Teachers should pay more attention to these students, asking them to remember and recite what they should master (vocabulary, sentence or grammar) and giving them more exercises to master the basic knowledge. For the students whose sores are between 90 and 100, they may lack some skills and techniques in dealing with the test. Thus, teachers had better teach them some skills or techniques to improve their scores.

b. The analysis of reliability

After the calculation by SPSS in the former study, the data can be collected in the following table.

TABLE 4.1
CASE PROCESSING SUMMARY

		N	%
Cases	Valid	30	100.0
	Excluded[a]	0	.0
	Total	30	100.0

a. Listwise deletion based on all variables in the procedure.

TABLE 4.2
RELIABILITY STATISTICS

Cronbach's Alpha	N of Items
.899	9

TABLE 4.3
ITEM-TOTAL STATISTICS

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Listening	54.42	391.50	.633	.890
MCQ	61.68	437.58	.723	.891
Cloze	64.22	444.36	.617	.895
Reading	51.88	324.98	.681	.904
Phrases Translation	65.38	453.58	.676	.897
Filling in the blank	66.27	439.37	.817	.890
Rewriting the sentences	64.6000	393.248	.837	.877
Sentence Translation	64.7000	348.666	.934	.863
Writing	60.5833	327.191	.848	.873

Theoretically, the range of Cronbach's α is between 0 and 1. If the α is higher, the relevance between different items will be better and the internal consistency of reliability will be higher. In the former study, the Cronbach's α is 0.899, which indicates that the consistency of scores is high and the test paper is of high reliability. Besides, a comparison between every single part and the whole test paper can be made in the above table. If the Cronbach's α (if item deleted) of some part is higher than 0.899, the items can be regarded as unreasonable items and they can be deleted. On the contrary, the items can be regarded as good items and can be stored in the test bank. Thus, among all parts in the test paper, only the value of reading comprehension is higher than 0.899, which indicates that there are some problems in reading comprehension. Maybe it is too easy or too difficult in some degree. It is no doubt that reading comprehension needs more improvement to correspond to the whole paper. Therefore, it is necessary to make a new study to probe into the reading comprehension and to find some problems in this section. So this study will give a detailed analysis of reading comprehension in the later part.

c. The analysis of validity

Construct validity can be established through the following measures: internal correlation, factor analysis and MTMM (multi-traits multi-methods). This paper is mainly focused on internal correlation. And 3 types of internal correlation can be listed as follows. Firstly, the correlation between different components should be low (0.3-0.5). Secondly, the correlation between two tasks in a testing component should be high (at least 0.5-0.7). Thirdly, each

© 2018 ACADEMY PUBLICATION

component should have a high correlation coefficient with the total score (above 0.7). So the results can be observed in Table 5.

According to Table 5, except for the Cloze (0.664), the correlation coefficient is all above 0.7, which meets the demands that each component should have a high correlation coefficient with the total score (above 0.7). And they are all significant at the 0.01 level with the total score. So it proves that the test is valid to some degree. Its results of this test are accurate and precise. Besides, as for the correlation of different tasks, almost the correlation of every two different parts is between 0.5 to 0.7, which is also reasonable. Sentence translation and writing have the highest correlation (0.892) with each other. In general, the test paper corresponds to the requirements of validity. That is to say, it is reasonable to test students' ability of English.

TABLE 5.
CORRELATIONS

	Listening	MCQ	Cloze	Reading	Phrases Translation	Filling in the blank	Rewriting	Sentence translation	Writing	Sum
Listening	1	.584**	.517**	.465**	.315	.519**	.585**	.649**	.556**	.723**
MCQ	.584**	1	.593**	.625**	.325	.584**	.588**	.668**	.620**	.759**
Cloze	.517**	.593**	1	.573**	.276	.529**	.460*	.546**	.495**	.664**
Reading	.465**	.625**	.573**	1	.475**	.507**	.556**	.720**	.586**	.805**
Phrases Translation	.315	.325	.276	.475**	1	.819**	.708**	.747**	.734**	.713**
Filling In the blank	.519**	.584**	.529**	.507**	.819**	1	.811**	.827**	.828**	.838**
Rewriting	.585**	.588**	.460*	.556**	.708**	.811**	1	.852**	.865**	.873**
Sentence translation	.649**	.668**	.546**	.720**	.747**	.827**	.852**	1	.892**	.955**
Writing	.556**	.620**	.495**	.586**	.734**	.828**	.865**	.892**	1	.900**
Sum	.723**	.759**	.664**	.805**	.713**	.838**	.873**	.955**	.900**	1

**. CORRELATION IS SIGNIFICANT AT THE 0.01 LEVEL (2-TAILED).
*. CORRELATION IS SIGNIFICANT AT THE 0.05 LEVEL (2-TAILED).

B. The Analysis of Reading Comprehension in This Study

In the previous section, the summary of the analysis of whole paper has been made clearly. And according to the analysis of the former study, the data show that the reliability of Reading Comprehension in the whole test paper is a bit unreasonable. Maybe there exist some problems in it. So it is extremely necessary to choose it as a single part with a detailed analysis in this study, which includes Facility Value (F.V.), Discrimination Index (D.I.), Reliability and Validity. There are totally 15 items in this part.

a. The analysis of F.V. and D.I. of Reading Comprehension

After collecting the data and calculating, F.V. and D.I. of Reading Comprehension can be listed as follows.

TABLE 6.
F.V. AND D.I. OF READING COMPREHENSION

item	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55
F.V.	0.93	0.97	0.60	0.90	0.47	0.80	0.67	0.40	0.47	0.13	0.37	0.80	0.27	0.73	0.13
D.I.	-0.12	-0.1	0.30	0.10	0.80	0.20	0.40	0.90	0.90	0.20	0.60	0.40	0.50	0	0.10

Based on the theory of F.V., the range is between 0 and 1. If the F.V. is higher, the item will be easier. However, the ideal range of the items is between 0.3 and 0.7. Then, observed from the above table, only 1/3 of the items are in the ideal range. Most of them are higher than 0.7. So these items tend to be too easy in reading comprehension. The lowest F.V. of those items is No. 50 and No.55 (equally 0.13), lower than 0.3. So these two items will be taken for example in further analysis. Theoretically, the range of D.I. is between -1 and 1. However, the ideal range is above 0.35. The higher, the better. According to the formula, D.I.=(H-L)/No. of H or L. So the data of every item can be listed in the above table. From the table, the D.I. of most items is below 0.35, which also indicates that the items are too easy to test their reading ability for students. Observing from the test paper, some problems can be found in reading comprehension. For example, the answers of most items can be found directly in the reading context. To some degree, this type of question is no significance either for teachers or for students. On one hand, teachers cannot find students' shortcomings from their answers because they are too easy to discriminate different levels of the items and to test students' real ability. On the other hand, students cannot find their own weakness of reading comprehension. If so, students will not improve their ability and make great progress in reading. Among all the items, D.I. of the item 48 and 49 is the highest (0.9). The following section will take them for example to analyze concretely.

Item 48 If you want to take the tour to Taiwan, you can book it by phone at_____

A. 4:00 p.m on Wednesday

© 2018 ACADEMY PUBLICATION

B. 4:30 p.m on Sunday
C. 8:00 a.m on Saturday
D. 8:00 a.m on Wednesday

Item No	Group	Options			
	(N=10)	A*	B	C	D
48	Top	9	0	0	1
	Middle	5	1	2	2
	Bottom	0	1	6	3

In item 48, F,V. is 0.4. D.I. is 0.9. The results show that the item has a high discrimination. The text is an advertisement on tourism. So this item is mainly to test students' ability to search information from an advertisement. Only if they find the office hours, they will choose the correct answer. Maybe some students do not understand the meaning of the stem. Maybe some students just guess the answer at random. This item is closely related to life. In general, this item is a good item.

Item 49 What does the phrase "*crumpled up*" mean_____

A. 撕碎 (tear up)
B. 弄皱 (crumple)
C. 揉碎 (triturate)
D. 展开 (unfold)

Item No.	Group	Options			
	(N=10)	A	B*	C	D
49	Top	0	10	0	0
	Middle	2	5	0	3
	Bottom	2	1	1	6

In item 49, F.V. is 0.47. D.I. is 0.90. This item is mainly to test the knowledge of vocabulary. And the options are easy to mix with each other. If students remember the vocabulary, they will choose the correct answer easily. Also some students can infer from the text or guess the meaning. But it is a good item to test students' ability of using the vocabulary.

Item 50 The speaker crumpled the bill up because he wanted to_____

A. Let more listeners want the bill
B. Make the listeners feel sorry for the bill
C. Make the bill look worse
D. Show the listeners that the bill was worthless

Item No	Group	Options			
	(N=10)	A	B	C*	D
50	Top	0	0	2	8
	Middle	2	3	3	2
	Bottom	5	2	1	2

In item 50, F.V. is 0.13. D.I. is 0.20. Most students think it difficult to make a choice. The key point is to read the text carefully and the following text has cues for this option. The difficulty is that option D has something ambitious. The word "*worthless*"occurs in the later text, so most students misunderstand it and consider it as the correct answer.

Item 55 Which of the following is not true according to the passage?

A. Diana Nyad became the first person to swim from Cuba to Florida.
B. President Barack Obama congratulated Diana Nyad on her success.
C. Diana Nyad began trying to swim from Cuba to Forida in 1978.
D. Diana Nyad faced many challenges during the swim.

Item No	Group	Options			
	(N=10)	A*	B	C	D
55	Top	2	3	3	2
	Middle	2	3	5	0
	Bottom	2	3	3	2

In item 55, F.V. is 0.13. D.I. is 0.10. It indicates that most students do not choose a correct answer. There are 2 key points in this item. Firstly, the key word "*not true*" in the stem of the item. Secondly, the option A has been easily misled as the true answer according to the text. Because in the original text, it says"*Nyad made history by becoming the first person to swim from Cuba to Florida without using a shark cage*". So many students make a mistake in this item because they do not read the original text carefully, leading to neglecting the phrase "*without using a shark cage*" in the original text. Practically, this item is not difficult, which is decided on whether students are careful or not.

© 2018 ACADEMY PUBLICATION

b. The reliability of Reading Comprehension

In this part, the reliability of Reading Comprehension calculated by SPSS can be concluded as follows. The valid data are 30. Table 7.2 shows that Cronbach's α of reading comprehension is 0.715. As a whole, it is not too low, which corresponds to the requirement of Cronbach's α. Thus, the result shows that the items of reading comprehension are relatively reasonable and reliable. Besides, Table 7.3 presents the reliability of every item in Reading Comprehension. According to the results in the table, the values of item 41, 42, 44 and 54 are higher than 0.715. So these items are not reasonably set up in the test paper and they can be deleted. But the rest of the items in Reading Comprehension are lower than 0.715, they can be thought as the good items to test students' reading ability and can be included in the test bank.

TABLE 7.1
CASE PROCESSING SUMMARY

		N	%
Cases	Valid	30	100.0
	Excluded[a]	0	.0
	Total	30	100.0

a. Listwise deletion based on all variables in the procedure.

TABLE 7.2
RELIABILITY STATISTICS

Cronbach's Alpha	N of Items
.715	15

TABLE 7.3
ITEM-TOTAL STATISTICS

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
N41	7.70	8.010	-.077	.730
N42	7.67	7.816	.113	.717
N43	8.03	6.723	.384	.693
N44	7.73	7.789	.049	.724
N45	8.17	6.626	.414	.689
N46	7.83	7.316	.219	.712
N47	7.97	7.068	.262	.709
N48	8.23	6.185	.618	.659
N49	8.17	6.144	.621	.658
N50	8.50	7.431	.220	.711
N51	8.27	6.547	.469	.681
N52	7.83	6.902	.419	.690
N53	8.37	6.861	.382	.693
N54	7.90	7.334	.176	.718
N55	8.50	7.431	.220	.711

c. The validity of Reading Comprehension

According to the basic theory of construct validity, the correlation between two tasks in a testing component should be high(at least 0.5-0.7) (Wei, 2007). From Table 8, the relevance of every two items cannot be up to the above standard, most of them are below 0.5. Item 41 is of the highest relevance with item 42 (0.695), whose correlation is significant at the 0.01 level. Therefore, Reading Comprehension, as a complete part in this test paper, is not valid enough. The validity also reflects the quality of Reading Comprehension. So this test paper should be further revised to be up to the standards of validity. Actually, one of the major problem in reading comprehension is that the items are too easy for students, which cannot test the real ability of students. From this point, teachers should choose various types of texts and set up different kinds of items to make them more difficult in line with the real ability of students. Another problem is that some students just guess answers randomly. So their attitudes towards the test should be paid more attention. Maybe their attitudes also have an effect on the validity.

© 2018 ACADEMY PUBLICATION

TABLE 8.
CORRELATIONS

	N41	N42	N43	N44	N45	N46	N47	N48	N49	N50	N51	N52	N53	N54	N55	T4
N41	1	.695**	.055	.356	-.018	-.134	-.189	-.327	-.286	.105	-.351	-.134	.161	.141	.105	.016
N42	.695**	1	.227	.557**	.174	-.093	-.131	-.227	-.199	.073	-.244	-.093	.112	.308	.073	.179
N43	.055	.227	1	.181	.218	.272	-.144	.250	.218	.320	.339	.102	.185	.277	-.080	.518**
N44	.356	.557**	.181	1	.312	-.167	-.236	.045	.089	.131	-.208	-.167	-.050	.050	-.196	.121
N45	-.018	.174	.218	.312	1	-.200	-.047	.464**	.464**	.026	.397*	.134	.342	.111	.026	.549**
N46	-.134	-.093	.272	-.167	-.200	1	.000	.238	.134	.196	.035	.375*	.113	.264	.196	.362*
N47	-.189	-.131	-.144	-.236	-.047	.000	1	.289	.378*	.069	.391*	.530**	.107	.053	.277	.452*
N48	-.327	-.227	.250	.045	.464**	.238	.289	1	.736**	.280	.508**	.238	.431*	.031	.080	.690**
N49	-.286	-.199	.218	.089	.464**	.134	.378*	.736**	1	.223	.536**	.301	.494**	-.191	.223	.694**
N50	.105	.073	.320	.131	.026	.196	.069	.280	.223	1	.109	-.049	-.015	-.207	.135	.331
N51	-.351	-.244	.339	-.208	.397*	.035	.391*	.508**	.536**	.109	1	.380*	.010	-.010	.312	.593**
N52	-.134	-.093	.102	-.167	.134	.375*	.530**	.238	.301	-.049	.380*	1	.113	.264	.196	.543**
N53	.161	.112	.185	-.050	.342	.113	.107	.431*	.494**	-.015	.010	.113	1	.193	-.015	.510**
N54	.141	.308	.277	.050	.111	.264	.053	.031	-.191	-.207	-.010	.264	.193	1	.015	.337
N55	.105	.073	-.080	-.196	.026	.196	.277	.080	.223	.135	.312	.196	-.015	.015	1	.331
T4	.016	.179	.518**	.121	.549**	.362*	.452*	.690**	.694**	.331	.593**	.543**	.510**	.337	.331	1

*. Correlation is significant at the 0.05 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).

V. CONCLUSIONS

This paper takes the reading comprehension in an English mid-term test as the object of the research, firstly accurately summarizing the former analysis of the whole test paper with the use of SPSS, including descriptive statistics, reliability and validity of the test paper. Then, based on the results of the previous study, it is found that some problems exist in reading comprehension. So Reading Comprehension is chosen as a single part to calculate its F.V., D.I., reliability and validity in this study. The results can be concluded as follows. From the point of F..V., D.I., reliability and validity of Reading Comprehension, the results show that some items of Reading Comprehension are not reasonable, which are too easy to test the real ability of students. Therefore, teachers should choose various types of texts and set up different kinds of items to make them more difficult to improve the quality of the test paper. Besides, SPSS, as an extremely necessary tool in language testing, plays an important role in language teaching. Based on the results of SPSS, teachers can exactly choose good items and delete bad items. And they can adjust different types of items to establish a test bank, which can improve the quality of the test paper and enhance the reliability and validity of a test. In brief, language testing has an extremely important influence on language teaching.

REFERENCES

[1] Xiaohong Cheng. (2015). The SPSS Analysis of Higher Vocational English, *Heihe Journal*, (9):99-101.
[2] Lyle F.Bachman & Adrian S.Palmer. (1996). Language Testing in Practice. Oxford: Oxford University Press.
[3] Lyle F.Bachman. (1999). Fundamental Considerations in Language Testing. Shanghai: Shanghai Foreign Language Education Press.
[4] Shen Zou & Weidong Dai. (2012). Language Testing (The Second Edition). Shanghai: Shanghai Foreign Language Education Press.
[5] Hongmei Wei. (2007). The Language Testing Analysis of Reliability and Validity of the Test Paper, *Journal of Foreign Art and Education Research*, (4):11-14.

Yingying Jin was born in Weifang, China in 1994. She is currently a master student in the University of Shanghai for Science and Technology, Shanghai, China. Her major is Foreign Linguistics and Applied Linguistics. Her research interests include syntax and foreign language teaching.

Xiaowen Qi was born in Nanjing, China in 1977. She received her PH.D. degree in linguistics from Shanghai International Studies University, Shanghai, China in 2015. She is currently an associate professor in the School of Foreign Languages, University of Shanghai for Science and Technology, Shanghai, China. Her research interests include syntax, foreign language teaching and cross-cultural communication.

© 2018 ACADEMY PUBLICATION