Advances in Social Science, Education and Humanities Research, volume 289
5th International Conference on Education, Language, Art and Inter-cultural Communication (ICELAIC 2018)

Implications of Multiple Test Forms of Chinese Matriculation English Tests

Juan Luo
College of Foreign Languages
Hunan University
Changsha, China

***Abstract*—The goal of this study was to discuss the impact of different forms of MET. This includes the analyses of test reliability, test difficulty, and item quality of three English tests. Results show that test forms differ in test quality, and then test equating was conducted with the equipcentile method. The implications of equated scores for college English teaching were discussed finally.**

Keywords—MET; comparability; college teaching

I. INTRODUCTION

The National Matriculation Test system in China (MET) is a high-stake standardized test which universities use to select new students. Since 2017, three multiple forms of MET designed by the Chinese Ministry of Education have been used nationwide and administered in different provinces.

However, this move has sparked debates concerning the comparability of test scores across test forms. Use of multiple forms of the same test is common in high-stakes tests, even if the scales measure the same construct, the raw scores on the different scales will not be comparable if the scales are not based on the same metric (Mâsse, Allen, Wilson, et al., 2006). Thus, the same raw score on the two different scales won't indicate the same level of the knowledge, the scores won't be comparable.

To compare scores from different forms is crucial for language testers, researchers, policy makers, and educators (Saida, Hattori, 2008). Score users need to compare the scores of test-takers who took different forms of the test, testing organizations need to report scores that are comparable and make test decision, educators need to place students into different levels based on comparable scores and adapt their teaching.

II. RESEARCH QUESTIONS

The solution to this problem is test equation, which is used to establish a mathematical relationship between two scales so that the scores are based on the same metric and thus are comparable (Kolen, Brennan, 1995).

This study uses English as the subject, and addresses the research questions:

- How does test quality compare between different test forms?
- How to equate scores of different test forms onto the same scale?
- What are the implications for college teaching?

III. METHOD

A. Participants

1157 senior students of high schools are determined as the research sample, and three tests forms of MET were given to them with one week interval, thus it is a common group equating design.

B. Test Forms

This research focus on three test papers, including Shanghai (2008), Shanghai (2009), and Jiangxi (2009), which are the most representative and similar in test content. For convenience, they are referred to as Test A, Test B and Test C. Equating of Test A and B represent the cross-year comparison within one province, while Test B and C represent the cross-province comparison within one year.

IV. DATA ANALYSIS

A. Analysis of Test Quality and Test Difficulty

The program IRTPRO has been used to estimate item parameters, which implements the method of Maximum Likelihood for item parameter estimation.

TABLE I. ITEM DIFFICULTY PARAMETER "B"

Test Form	Item Number	Range	Minimum	Maximum	Mean	Std. Deviation
A	84	3.48	-1.24	2.24	.5389	.85837
B	84	4.14	-1.63	2.51	.3729	.99814
C	85	6.06	-2.39	3.67	.9774	1.08706

TABLE II. ITEM DISCRIMINATION PARAMETER "A"

Test Form	Item Number	Range	Minimum	Maximum	Mean	Std. Deviation
A	84	.83	.06	.89	.4862	.19430
B	84	.83	-.17	.66	.2367	.17797
C	85	2.72	-.01	2.73	1.1705	.53256

"Table I" and "Table II" shows the descriptive analysis of item parameter estimates. Comparatively, the mean of *b* and *a* of Test C are substantially larger than those of Test A and B, suggesting the difficulty level of Test C is higher than the other two, meanwhile, it has high item discrimination.

B. Test Reliability Analysis

One fundamental question in evaluating test quality is how accurately a test has measured an individual's ability. In IRT, this question can be addressed by computing test information function (TIF for short).

"Table III" demonstrates that along the ability scale, the TIF of Test C is significantly the largest, and the expected S.E is the lowest. The TIF in "Table III" shows that the three tests are doing well in estimating ability, moreover, Test A and Test B are very close in test precision across the full ability range, while Test C yields a test with obviously most test precision. Furthermore, in the ability range [-2, 0], "Fig. 1" shows the relative efficiency of TIF, which implies that the amount of TIF of the three tests is:

Test C >Test A >Test B.

TABLE III. TIF FOR THE THREE TEST FORMS

Test Form	θ	-2.8	-2.4	-2	-1.6	-1.2	-0.8	-0.4	0	0.4	0.8
A	Test Information:	10.57	14.28	18.64	22.93	26.07	27.03	25.7	22.95	19.62	16.1
	Expected s.e.:	0.31	0.26	0.23	0.21	0.2	0.19	0.2	0.21	0.23	0.25
B	Test Information:	10.1	13.24	16.65	19.71	21.62	21.96	20.94	19.1	16.97	14.74
	Expected s.e.:	0.31	0.27	0.25	0.23	0.22	0.21	0.22	0.23	0.24	0.26
C	Test Information:	10.36	14.98	21.34	29.38	37.74	41.19	36.04	27.29	19.39	13.57
	Expected s.e.:	0.31	0.26	0.22	0.18	0.16	0.16	0.17	0.19	0.23	0.27

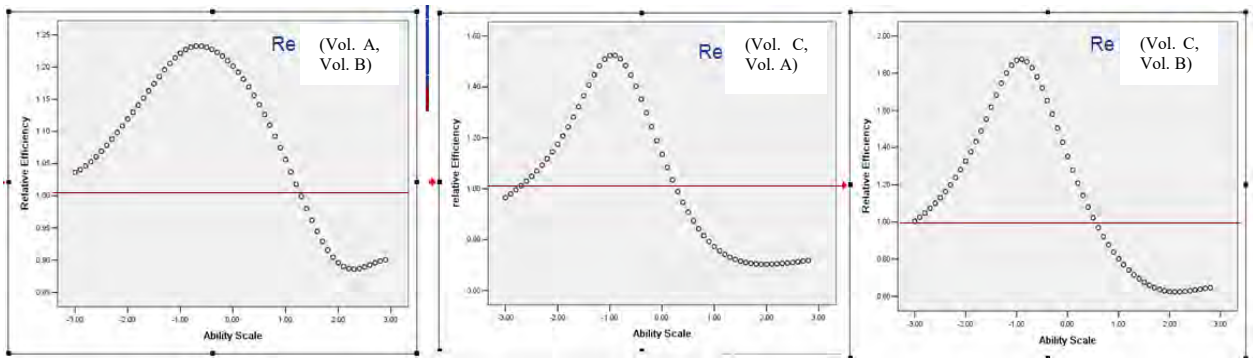

Fig. 1. Relative efficiency of TIF of three tests.

TABLE IV. EQUATION OUTCOME BY PERCENTILE METHOD (EXCERPTS)

Test A Score	Percentile Rank	Test B Score	Percentile Rank	Test C Score	Percentile Rank
89	42	89	37	89	50
95	50	95	42	95	62
100	58	100	47	100	70
102	62	102	50	102	73

C. Test Score Equation

To address the questions as "how to put the test score onto the same scale", the equating method of equipcentile is applied.

"Table IV" shows the same raw score on different tests has different percentiles in the testing group. Relatively, the same raw score on Test C has the highest percentile rank, while Test B has the lowest percentile rank, which suggests the same raw score doesn't reflect the same level of ability. This is due to the difference in test difficulty among the three tests, as Test C is the most difficult test, Test B is the easiest one, the same raw score reflects different abilities level in the same testing group. For an example, the score of 100 on Test A corresponds to the percentile rank of 58, while the same score on Test B and Test C ranks 47 and 70 in the same group respectively. It indicates that for different examinees with the same score on various test forms actually have very diverse abilities.

V. CONCLUSION

This study investigated test quality in terms of reliability and difficulty of three English tests of MET. The output shows that the three tests are not equivalent in level of difficulty and test accuracy. An equipcentile equating study has been carried out to put the raw scores of the three tests onto the same scale. The percentile distribution shows that the same raw score on different tests has different percentiles, which suggests the same raw score doesn't reflect the same level of ability due to the difference in test difficulty.

The findings indicated that there are some negative consequences on college teaching and learning. Due to the difference in test difficulty, students admitted nationwide and placed in the same level in the college had very diverse abilities, and there might be a huge difference between students in the same classes. It is harder to maintain a good teaching pace and meet the needs of all the students‑stronger students were demotivated, and weaker ones were struggling. The difference in the examinees' ability also leads us to question the fairness of admission decision, and whether the decision provides equal education opportunity for all the testing groups.

This study provides the solution to the above problems by equating the scores of various test forms and makes them comparable. For testing organizations, the equation outcome can server as a scientific reference, which helps them to report comparable scores across regions and years; for policy makers, it help enhance the fairness of the admission decision. For educators, students are placed into the same level according to their equivalent abilities, and the difference in their ability is reflected by the difference in the scaled score, which makes the educators more aware of how to adjust their teaching.

REFERENCES

[1] Kolen MJ, Brennan RL. (1995). Test Equating: Methods and Practices. New York: Springer-Verlag.

[2] Mâsse L C, Allen D, Wilson M, et al. (2006). Introducing equating methodologies to compare test scores from two different self-regulation scales. Health education research, 21(suppl 1): i110-i120.

[3] Saida C, Hattori T. (2008). Post-hoc IRT equating of previously administered English tests for comparison of test scores. Language Testing, 25(2): 187-210.