International Journal of New Developments in Engineering and Society
ISSN 2522-3488 Vol. 4, Issue 2: 70-75, DOI: 10.25236/IJNDES.040214

The Construction and Application of the Vertical Scale System of College English Grading Test

Lixia Wu

Yancheng Kindergarten Teachers College Preschool Education Department, Jiangsu Yancheng 224002, China

ABSTRACT. *It is an important purpose of testing to evaluate and recommend talents fairly. However, due to the lack of vertical measurement, the fairness of the assessment system of College English grading teaching has been questioned. At the same time, this kind of questioning has a negative backwash effect on the classified teaching mode. Therefore, it is of great theoretical and practical value to analyze how to construct the vertical scale and eliminate the drawbacks of College English graded teaching evaluation system, so as to realize the fairness of the evaluation system and the teaching concept of teaching students according to their aptitude.*

KEYWORDS: *Vertical scale, College english graded teaching, Disadvantages of evaluation system*

1. Introduction

In recent years, in response to the call of College English reform, colleges and universities all over the country have launched the College English Graded Teaching Mode in full swing. However, there are some failed colleges and universities. Many of the reasons for the failure are due to the lack of an effective test evaluation system. For students of different levels, if the traditional “one size fits all” test method is still used, the effectiveness and fairness of the measurement results are obviously not convincing. In view of this problem, the introduction of the vertical scale method in the field of psychometrics and the establishment of the vertical scale system can provide ideas to solve the problem of test and evaluation in College English teaching.

2. Connotation and Characteristics of Vertical Scale

Vertical scale, also known as vertical equivalence or cross level scale, is to put the scores obtained by subjects of different levels in tests with the same construct and different difficulty into a common integrity scale, so that the progress of the subjects' stage in a specific subject test can be tracked and compared with the

ISSN 2522-3488 Vol. 4, Issue 2: 70-75, DOI: 10.25236/IJNDES.040214

implementers. Holland and dorans call the process of converting the scores from the two tests into links, which can be divided into three categories: prediction, scale alignment and equating. Prediction is the earliest form of score linking. Its purpose is to minimize the expected error of score of dependent variable or standard variable from other prediction variables. The purpose of numerical correction is to convert the scores obtained in different tests to a common scale. Equivalence is the most strict requirement for tests to be linked. Its purpose is to establish a link between the two forms of tests, so that every The scores in each test can be regarded as the scores from the same test to meet the needs of practical application. The vertical scale belongs to the second category of numerical correction. When designing the vertical scale system, the basic approach is to make the adjacent groups have a common test item, which is called anchor question, and then establish a common scale based on it. There are two ways to build a common scale: the classical test theory (CCT) and the item response theory (IRT). If the IRT method is selected, the second problem to be considered is the selection of the calibration method of the scale: the first method is to calibrate the items, personnel parameters and common items of the test form at the same time; the second method is to calibrate the parameters of different test forms separately, and then use Some numerical linking method puts them on the same scale. Other issues to be considered include: the length of the common project set, the selection of the base year and the selected computer software, etc [1].

3. The Present Situation of College English Graded Teaching

In 2003 and 2007, China's Ministry of education successively launched the relevant documents of College English teaching reform, College English curriculum teaching requirements (Trial) and College English curriculum teaching requirements. According to the different English basic conditions when students enter the University, three levels of requirements are put forward for College English teaching, namely, general requirements, higher requirements and higher requirements, classified guidance and Teaching students according to their aptitude is the core of the reform. In the context of this kind of reform, graded teaching is the concrete reform measure of College English teaching. It has been more than 10 years since the earliest reform was implemented. According to the requirements of the documents, most schools in the country have begun to implement the graded teaching mode. However, due to different situations in different regions, the implementation of the specific situation also has its own characteristics. At present, the focus of controversy mainly lies in the test and evaluation system which tests students' learning situation after grading. At the beginning of the reform, many colleges and universities are still in a state of confusion. For convenience, some colleges and universities still use the same test for students of different levels in the final test, which inevitably makes people question the fairness of the evaluation. Schools that have been implementing graded teaching for a long time have begun to design papers of different difficulty for different levels to improve the fairness of the examination. However, there are also corresponding problems: how to compare the scores of students of different levels in different difficulty papers? At present, in our country, the examination scores of many schools directly affect the students' vital

ISSN 2522-3488 Vol. 4, Issue 2: 70-75, DOI: 10.25236/IJNDES.040214

interests, such as the evaluation of awards, joining the party and applying for studying abroad in the future. Therefore, it is urgent to establish an objective, scientific and effective test evaluation system [2].

4. The Construction of the Vertical Scale System of College English Grading Test

4.1 Preparation of Two Way Test Schedule

The preparation of the test bi-directional detail table is the first step of the test vertical scale. In order to make the test scores comparable and interpretable, it is necessary to make clear the contents, objectives and the correlation between them in the process of compiling the two-way detailed list of each level. This process will directly affect the reliability and validity of the test, and become the key to the success of the vertical scale [3].

4.2 Vertical Scale Design

Vertical scale design, that is, which way to obtain test data, makes the unified scale system to be established more reasonable and stable. In general, the anchor question unequal group design is the main way to construct the link relationship of tests with the same idea and different levels. Moreover, when the learning content gap between different levels is smaller and the information quantity is less attenuated, the “link” strength of anchor questions between tests is greater, and the stability of vertical scale is stronger. Therefore, using the double anchor test to establish the link between adjacent level tests first, and finally realize the vertical scale of all levels of tests through the link between adjacent levels, will ensure the link function of anchor test to the greatest extent.

4.3 Anchor Test Design

Anchor test design, also known as common question design, is to include a part of the same questions in two levels of tests, so as to establish the link between different levels. According to the characteristics of grading test, the “internal anchor design” is more feasible. As for the proportion of anchor questions, it is about 1 / 3 or 1 / 4 of the test length to ensure the validity of the equivalent data. As for the type of anchor question, the research shows that: when the anchor question is purely objective, the error of equivalence is the smallest; when the anchor question is purely subjective, the error of equivalence is the largest; when the mixed question is used as anchor question, the error of equivalence is between the two. Therefore, the double anchor design with built-in and objective questions becomes the final choice of anchor test [4].

ISSN 2522-3488 Vol. 4, Issue 2: 70-75, DOI: 10.25236/IJNDES.040214

4.4 Selection of Equivalent Data Conversion Method

As for the conversion methods of equivalent data, the common ones are: the equivalent data conversion method based on CTT, Thurstone absolute scale method and the equivalent data conversion method based on IRT. CTT equivalence method has a strong dependence on the sample of transformation relationship, and the uniqueness of equivalence conditions is difficult to meet, so it can not ensure the symmetry and fairness of transformation relationship. Thurstone absolute scale method is based on the assumption that the ability of subjects is in normal distribution. Therefore, when the ability of subjects is in negative distribution, in order to obtain the normal distribution, there will be the problem of scale "expansion" in the high section. The IRT equivalence method uses the potential ability value estimated according to the answer pattern of the subjects. It does not presuppose the ability distribution of the subjects in advance. It is the actual reflection of the ability distribution of the subjects, and the ability parameters of the subjects are invariant, so it has the advantages that other theories cannot reach. So far, the IRT equivalence method has become the main method of constructing the vertical scale system and an indispensable part of the construction of the question bank [5].

4.5 Estimation Method of Project Parameters

The estimation of project parameters involves the question of which IRT model to adopt. Considering the convenience of calculation, logistic model is often used. As the 1plm model requires that the discrimination degree of each topic is equal and equal to 1, this requirement is difficult to achieve, and the cost is high; at the same time, the project parameters in 3PLM model are difficult to accurately estimate, that is, the project feature surface is gentle, unless a good initial value is taken, the logarithmic likelihood function is difficult to reach the extreme point. Moreover, for 3PLM, there is no sufficient statistic for capability parameters. So 2PLM model is the best choice [6].

4.6 Estimation Method of Equivalent Coefficient

Under the framework of item response theory, the estimated values of item parameters and ability parameters from different tests are usually not on a unified scale. Mathematically speaking, in order to make the estimated parameters of the two tests comparable, it is necessary to transform the measurement system, that is, to estimate the equivalent coefficient. There are many methods to get the equivalent coefficient under IRT framework, including parameter equivalent method and project characteristic curve equivalent method. The equivalent method of project characteristic curve is a typical equivalent method of project response theory, which is more complex, but the result of equivalent is more reliable. At present, the popular stocking Lord method (SL) and Haebara method (H) belong to this equivalence method. The two methods have their own advantages and disadvantages. The

ISSN 2522-3488 Vol. 4, Issue 2: 70-75, DOI: 10.25236/IJNDES.040214

behavior of SL method is: when the equivalent coefficient a is less than or equal to 1.0, SL method performs better, while h method performs better when a is more than 1.0; when the random error value is large, SL method performs better in a larger range, while when the random error is small and the random error value is medium, the results of the two methods are consistent. Therefore, College English grading test can choose a more reliable method based on the test results [7].

4.7 Presentation of Equivalent Results

How to report the test results to the students, and make the report results conducive to understanding and interpretation, is also an important content that must be considered in the construction of vertical scale system. At present, the scores of test reports can be in the form of pass rate, standard score, percentile level and other forms. Different vertical equivalence methods will use different forms of score report, but which is more reasonable, researchers have no unified conclusion. Petersen suggested that the score report form used should be conducive to the interpretation of the score meaning and minimize the possibility of the score being misunderstood. This is also the theoretical guidance for the selection of presentation mode of equivalent results in this study. Based on the above research, in the process of constructing the vertical scale system of CET, researchers can choose the methods with relatively accurate results, relatively convenient operation and relatively easy for students to understand according to the actual situation. In this way, it can not only effectively eliminate the drawbacks of the College English grading teaching evaluation system, but also effectively improve the feasibility of the construction of the test vertical scale system [8].

5. Conclusion

In order to achieve the principle of fairness, it is necessary to build a vertical scale system in College English Grading Test, and the specific theoretical guidelines and previous classic cases have proved the feasibility of this approach. In the future, more research needs to be devoted to the specific evaluation methods of colleges and universities, and the ways to cultivate talents in the field of English teaching who can closely link the knowledge of psychometrics and College English teaching.

References

[1] Chen Li (2014). Analysis of the drawbacks of the evaluation system of College English Graded Teaching by vertical scale [J]. Journal of Xi'an Foreign Studies University, vol. 22, no. 2, pp. 76-78.

[2] Min Shangchao, He Lianzhen (2016). Construction of English listening development scale -- Application of IRT vertical equivalence [J]. Chinese foreign language, vol. 13, no. 4, pp. 70-77.

[3] Xiao Yunan, Luo Juan (2019). Cognitive diagnosis of listening comprehension in College English Grading Test [J]. Journal of Hunan University (SOCIAL

ISSN 2522-3488 Vol. 4, Issue 2: 70-75, DOI: 10.25236/IJNDES.040214

SCIENCE EDITION), vol. 33, no. 1, pp. 113-118.
[4] Wang long (2017). Research on the backwash effect of College English entrance grading test results on College English Teaching -- Taking 2016 freshmen of Lanzhou University of technology as an example [J]. Journal of Jilin University of education, vol. 33, no. 3, pp. 76-787.
[5] He Lixin(2013). Construction of College English Grading Test Question Bank Based on project response theory [J]. JOURNAL OF SHENYANG NORMAL UNIVERSITY (SOCIAL SCIENCE EDITION), vol. 37, no. 5, pp. 78-80.
[6] Gan Ling, Jiang Yemei (2013). A study on the validity of reading comprehension content in College English final grading test -- a case study of different class tests [J]. Journal of Guangxi Normal University (PHILOSOPHY AND SOCIAL SCIENCES EDITION), vol. 49, no. 5, pp. 155-160.
[7] Zhang Jing (2008). Test and evaluation of College English Graded Teaching [J]. Journal of Ezhou University, vol. 15, no. 6, pp. 60-62, 80.
[8] Yang Jiang, Cao Xinyu (2019). Research on the construction of College Students' English hierarchical reading intelligent library [J]. Science and education guide - Electronic Edition (late ten days), vol.12, no.8, pp. 184-187.