**Abstract**

The aim of this study is to introduce the concept of equating, its implications and methods in the domain of measurement and evaluation. In this paper the two kinds of classic equating and item response theory (IRT), through collecting data and analyzing them were investigated along with considering advantages and disadvantages of these methods. According to these different cases errors of equating were reckoned through computer software’s BILOG and SPSS. Considering a series of available data and methods of measurement, and using equating, the test were completely explained through the data drawn from two tests. Consequently, the two tests were equated with the same scale, and accuracy of equating was estimated. Finally, the advantages and potential applications of equating of tests were considered in classic theory and item response theory. This study has implications for educational measurement and testing procedures especially for adaptive testing and test constructing.

© 2011 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of Dr Zafer Bekirogullari.

*Keywords*: Equating; linking; scaling; calibration; measurement; classic theory; Item Response Theory (IRT); anchor item; difficulty; discrimination;

**1. Introduction**

Equating is a statistical method to relate two or more tests, or to base two or more tests on a same scale, which is called ‘equating of test’ (Hambleton & Swaminathan, 1985). Other terms like linking (Vale, 1986), calibration (Wright, 1968), and scaling (Hambleton, Swaminathan & Rogers, 1991) were also used to describe methods of
equating. Some researchers (for example, Kolen & Brennan, 1995), however, argue that a method can be called equating when it is used to equate two forms of test with the same content, and the other similar method should be called scaling and linking. In this paper, with some simplification, the term of equating (which means establishing some relations between the tests) is used, for it is very hard to determine whether the content and difficulty of two tests to be equated are exactly the same.

Methods of equating are called differently, but in general they can be categorized in this way: horizontal and vertical (Cook & Eignor, 1991). Horizontal equating is proper when for the security of test several forms of test are needed. These forms are not the same, but it is expected that they are similar in their content and difficulty. When the difficulty, reliability, and content of tests are so different from one form to another, few methods of equating can properly work (Cook & Eignor, 1991). Furthermore, it is expected that testinees, who are subject to the forms of test, have rather equally distributed abilities. When there is so much difference in distribution of abilities between testinees, traditional methods of equating (like linear equating and equipercentile equating) do not response (Kolen & Bernnan, 2004).

On the other hand, vertical equating is an attempt to equate the grades of two tests which are deliberately set with different levels of difficulty, and yet measuring the same general realm of knowledge or one skill-based area. Moreover, as opposed to horizontal equating, in this kind of equating the distribution of abilities between testinees is different from one level to another. Obviously the problem of vertical equating is significantly more complicated than that of horizontal equating. Some of specialists of measuring maintain that vertical equating should not be considered in equating of tests, for the content of tests is often different from one level to another (Kolen, 1988).

Some of the other specialists of measuring, however, argue that determining that whether tests can be equated should be dependent on the possibility of establishing the equating conditions. It is generally accepted that before equating the tests four conditions are needed: (a) similar ability, (b) equity, (c) population invariance, and (d) symmetry (including: Angoff (1984), Kolen & Brennan (1995), Lord (1980), Petersen, Kolen & Hoover (1989)). Similar ability means that, to be equated, tests should measure similar abilities (or qualities, characteristics, and skills).

The basic problem of this study is to introduce the concept of equating, its implications and methods in the domain of measurement and evaluation. In this paper the two kinds of classic equating and item response theory (IRT), through collecting data and analyzing them were investigated along with considering advantages and disadvantages of these methods.

2. Theoretical Background

The reasons for equating of test are different from one situation to another, but in general they can be classified in three categories: one important and fundamental reason for equating can be improving the cohesion and integrity of grades of test in order to not being forced to testine several times. To assure one of justice and equity of a test, or to neutralize the effect of exercise, in the process of setting the test several forms are often needed. For measuring the learning merit of children, for example, dual forms were set and published in a learning merit test (Michaelides, 2004).

The second reason for equating is that through this the grades of test become changeable. While several tests are used to measure the same variables or qualities, the grades of these tests cannot often be used in place of each other, because they are based on different scales. In order to compare testinees, or equivalent criteria series, in the test, first it is necessary to base the grades of test on the same scale through equating of test (Mislevy, 1992).

The third reason is continuity of test. Multiple tests are used to measure the growth or change of an ability or quality in different levels of ability. In this situation, tests are often deliberately set with different levels of difficulty. In order to measure this growth or change in ability, we can set different tests in different forms with different levels of difficulty. Thus, when the difficulty of test is changed, the ability of measuring growth or change cannot be guarantied. Therefore, these tests should be again based on the same scale. According to two first reasons, equating of test is categorized under the name of horizontal equating, and according to the last reason is called vertical equating. (For more information concerning the reasons for equating of test, see Mislevy (1992), and Peterson and others (1989)).
2-1- Designs of equating

Classical methods of equating were described in detail by Angoff (1971) and Kolen (1988). In general, the methods fall into the two main categories: equipercentile equating and linear equating. Equipercentile equating is accomplished by considering the scores on tests X and Y to be equivalent if their respective percentile ranks in any given group are equal. Strictly speaking, in order to equate scores on two tests, the tests must be given to the same group of examinees. In practice, the process typically is carried out by giving the tests to randomly equivalent groups of examinees.

Like other aspects of setting tests, equating begins with collecting the data. Generally four designs are used to collect the data: (a) single-group design, (b) equivalent-groups, and (c) anchor-test design (d) Common-Person Design. In single-group design two or more forms of test are presented to the same group of testees. The advantage of this design is that the errors of measuring will be reduced. The reason of this reduction is that just one group of testees enter in the equating process of test. Therefore, the differences between tests cannot be confused with those between the groups. Since in this design it is necessary to testees to response the items of several tests, fatigue is a basic problem, especially when physical and intellectual activities are included. The effect of exercise is another problem that should be considered. If awareness of a test improves testeesing performance, so the test that is to be presented later would seem easier. In order to avoid fatigue and the effects of exercise, some kinds of spiraling process should be used. For example, succession of presenting the forms of tests can work as a modifying factor (Kolen, 1988, Kolen and Brennan, 1995).

In equivalent groups design, two equivalent tests are presented to two equivalent groups of testees. The groups could be randomly selected, and for this it is sometimes called random groups design. The advantage of this design is that the problems of single-group design, like the effects of fatigue and exercise, can be reduced. Moreover minimum time is needed and the test can be completed in one presentation. The defect of this design, however, is that unknown levels of taking sides are entered in equating process, for the groups are not often exactly the same in distribution of their abilities. In order to control the taking sides, greater samples are often needed in this design (Kolen, 1988, Kolen and Brennan, 1995).

In anchor-test design, which is also called common-Item nonequivalent groups, tests are presented to two different groups of testees. As opposed to equivalent groups design, the groups can be different in their distribution of abilities. In addition, a series of common items like anchors are included in two tests. Finally, Common-Person Design: The two tests of be linked are given to two groups of examination, with a common group of examination taking both tests. Because the testing will be lengthy for the common group. This design has the same drawbacks as the single-group design.

3. Methodology

3-1- Sample size

To answer the intended items, the subjects for this population were considered. The people were all Iranian and included all the 2010 who were taking the test in the field of Test of language by The Iranian measurement organization (TOLIMO).

The sample size with the sampling error of 0.5 % and confidence level of 95 % includes 1054 participants for form A and 1241 participants for form B;

3-2- Instrumentation

Test of language by The Iranian measurement organization (TOLIMO) was held in two steps in May and July of 2010. This test is based on three sub-scales of vocabulary, structure, and listening. In this test eight
questions were selected as anchor questions. Grades of the forms of A and B were compared according to the classic method and IRT.

4. Results

<table>
<thead>
<tr>
<th>Table 1: Deceptive statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>form</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
</tbody>
</table>

The results driven through the classic method contained some problems. If the test is really parallel and with proper validity, the classic method can be used. The results showed that in the classic method there is a little equating between the forms of A and B.

In the classic method two conditions of symmetry and invariance should necessarily be regarded. In the method of regression for determining the constants, which is used in equating formula, it is not needed to regard symmetry. The necessity of invariance comes from the fact that the method of equating should be independent from the sample group. These conditions are generally very difficult to be established in the classic method.

The results of the analysis in the classic method showed that the regression indexes were influenced when respectively the forms of A and B were the references, in a way that with the results driven from the first state was significantly different from the results of the second one. The results showed that in the classic method equating is dependent to the observed sample, insofar as the observed sample being changed, the mean and standard division from the two samples for a test were significantly different. IRT can resolve two limits of classic method, symmetry and invariance.

In IRT if the model be fitted with the data, parameters of ability can be measured. In this case, parameters of difficulty ($b_A$) and discrimination parameter ($a_A$) in the form of A and difficulty parameter ($b_B$) and discrimination parameter ($a_B$) in the form of B can be compared. In addition, with IRT the parameter of ability (TETA) in the form of A and the parameter of ability in the form of B can be compared. The results showed that the parameter of discrimination and difficulty in different mentioned states were not significantly different. The measured parameter of ability in two groups also did not show a significant difference.

Acknowledgment

I am heartily thankful to professor Delavar, whose encouragement and guidance from the initial to the final level enabled me to develop an understanding of the subject.
References:


