Does the Rasch Model work for Equating?

آیا مدل اندازه‌گیری راش برای همطرازسازی آزمونها کاربرد دارد؟

Rassoul Sadeghi
PhD student
University of New South Wales

Jim Tognolini, PhD
ACER General Manager,
Sidney Office

رسول صادقی
دانشجوی دکتری
دانشگاه جدید ولز جنوبی

دکتر جیم توگنولینی
مدیر کل سازمان پژوهشهای آموزشی، دفتر استرالیا

Abstract

The advent of modern psychometric theory, Item Response Theory (IRT), has enabled performance to be compared over time, across academic year levels where different tests (different items assessing the same construct) have been used for different student groups on different occasions. In order for this to occur, the tests have to be equated. Once they are equated, the students' performances can be represented along the same scale. Once they are on the same scale they can be directly compared e.g. comparison of Year 3, 5 and 7 in a subject and their performances can be compared to predetermined cut-scores. Test equating of this type is currently used widely in Australia to identify the percentage of students to be 'at risk' (below benchmark). The results from two equating procedures (relative anchoring and concurrent equating) used with Rasch (1960) measurement models are compared, as fit to the model gets progressively worse. The research question is what happens to students' marks as fit to the model varies? Data in this study were generated from the one-parameter logistic model using the Simulation Program for Rasch Data (RUMMSims). The findings of the present study indicate that when data fit the Rasch model there is no significant difference between results produced from the different equating procedures. However, as data fit to the model gets progressively worse, the equating results that emerge from applying different equating procedures generate significant variations.

Key words: rasch model, equating, item-response theory, psychometric

Correspondence concerning this article should be addressed to Rassoul Sadeghi, PhD student. The university of New South Wales, Sydney, Australia.

چکیده

پیدایش نظریه‌های جدید روانسنجی، از جمله نظریهٔ سؤال-پاسخ این امکان را فراهم آورده که بتوان عملکرد دانش‌آموزانی را که در زمانهای مختلف با آزمونهای متفاوت سنجیده شده‌اند (با این فرض که آزمونهای مذکور خصیصه مشترکی را می‌سنجند) با یکدیگر مقایسه کرد. به منظور این مقایسه، آزمونهای مذکور باید همطراز شوند. فرآیند همطرازسازی موجب می‌شود عملکرد دانش‌آموزان مختلف در آزمونهای گوناگون در یک مقیاس واحد بیان و مقایسه گردند. این نوع همطرازسازی به طور گسترده‌ای در کشور استرالیا به کار گرفته می‌شود تا درصد دانش‌آموزان در معرض خطر آموزشی را شناسایی و برایشان برنامه‌های جبرانی ارائه کنند. پژوهش حاضر نتایج حاصل از دو روش همطراز سازی (نسبی و همزمان) با استفاده از مدل اندازه‌گیری راش (1960) را با یکدیگر مقایسه می‌کند و به دنبال پاسخگویی به این سؤال است: اگر داده‌های گردآوری شده برای همطرازسازی با مدل راش تطابق نداشته باشند یا تطابق این داده‌ها با مدل راش بتدریج کمتر شود نمره‌های حاصل از همطرازسازی مذکور دستخوش چه تغییراتی می‌شوند؟ به منظور شبیه‌سازی داده‌ها با میزان تطابق متفاوت با مدل راش از برنامه شبیه‌سازی کامپیوتری (RUMMSims) استفاده شد. یافته‌ها نشان می‌دهند که وقتی داده‌ها با مدل راش تطابق دارند نتایج حاصل از دو روش همطرازسازی با یکدیگر قابل مقایسه هستند، اما هرچه تطابق بین داده‌ها و مدل راش کاهش می‌یابد نتایج حاصل از دو روش همطرازسازی به تغییرپذیریهای معناداری منتهی می‌شوند.

واژه‌های کلیدی: مدل راش، همطرازسازی، نظریهٔ سؤال-پاسخ، روانسنجی.

Introduction

In many testing situations it is frequently necessary and desirable to have several forms of a test for a variety of reasons, such as maintaining test security and enabling an individual to take a test more than once. Multiple forms are also essential for situations such as educational and vocational admission testing and longitudinal studies such as those that try to monitor developmental trends in children's cognitive, social and emotional abilities. When tests are used in these situations, the test forms should be equated onto a common metric to convert the raw scores obtained from two different forms of a test "so that scores derived from the two forms after conversion will be directly equivalent" (Angoff, 1971, p. 562). In the case of college and vocational selection, for example, equating is essential because comparisons are made between persons who sit for different forms of the test; without equating, persons who take the more difficult form would be at a disadvantage relative to persons who take the easier form.

The process of transforming scores on one test so that they can be compared directly to scores on another is referred to as equating, scaling or linking. As an integral part of the test construction process, test equating has received widespread coverage in the measurement literature. It is beyond the scope of this paper to consider all aspects of test equating. However, it is the aim of this paper to provide an overview of some of the more common methods for equating different test forms.

Equating, Scaling and Linking tests

Equating, scaling and linking are terms used to describe the empirical procedures used in transforming the scores of tests to ensure that it makes no difference, which set of items students have taken. After equating has been carried out, it is possible to compare the performance of students, even though the students have scores based upon tests composed of different items.

Beguin (2000) makes the following distinction between the terms equating, linking and scaling which are the terms that describe the statistical procedures used to adjust the scores on different test forms so that they can be used interchangeably (see Angoff, 1971; Kolen and Brennan, 1995).

Equating is the process used to adjust the scores on equivalent test forms. A process related to equating but different in purpose is linking… Linking is used for tests that are purposefully built to be different in statistical characteristics. From a statistical point of view, equating is a special case of linking or scaling to achieve comparability (Beguin, 2000, page 3).

In essence, equating measures ensures that the measures are interchangeable. Scaling on the other hand refers to the process of associating numbers with the performance of students. When two tests have been equated, they are placed on the same scale. However, when two tests have been scaled they have not necessarily been equated (Kolen, 1985)

A definition of equating promulgated by Angoff (1971) states that to equate two test forms is

… to convert the system of units of one form to

the system of units of the other – so that scores derived from the two forms after conversion will be directly equivalent (Angoff, 1971, page 562).

Lord (1977, 1980) has proposed a definition of equating which introduces the notion of equity.

Tests X and Y can be considered to be equated if and only if it is a matter of indifference to each examinee whether he takes Test X or Test Y (Lord, 1977, page 128).

There are a number of implicit conditions that should be met before equating test forms. These conditions are summarised by Peterson, Kolen & Hoover (1989) as follow:

1) The tests being equated must measure the same variable- unidimensionality. For example, an analogy from the physical sciences would be equating degrees Fahrenheit and degrees Centigrade. Both are measures of temperatures. Similarly the equating of different currencies, such as Australian dollars, French francs and Italian lira is possible because they are measures of the same variable, purchasing power. In the case of equating test, it makes sense to equate tests that obviously measure the same variable. For example, equating a reading literacy test from Western Australia to one from Victoria is worthwhile. Angoff (1971) would suggest, however, that it makes little sense to equate tests measuring performance on different variables. For example, he suggests that equating a test that is a measure of arithmetic achievement to a test, which is a measure of artistic aptitude, is worthless.

2) The transformation is the same regardless of the group from which it is derived- population invariant. In other word, the resulting equivalence should not depend on the students, whose responses are used to develop the transformation, thus making the equating generalisable.

As Angoff (1971) states:

…in order to be truly a transformation of only systems of units, the conversion must be unique, except for random error associated with the unreliability of the data and the method used for determining the transformation; the resulting conversion should be independent of the individuals from whom the data were drawn to develop the conversion and should be freely applicable to all situations (Angoff, 1971, page 562).

The condition is an extension of a basic measurement principle developed explicitly by Thurstone (1959).

If a scale value is to be regarded as valid, the scale values of the statements should not be affected by the opinions of the people who help to construct it. This may turn out to be a severe test in practice, but the scaling method must stand the test before it can be accepted as being more than a description of the people who construct the scale (Thurstone, 1959, page 228).

Rasch (1960) also stressed the need for this kind of invariance and referred to it later in his writings as “specific objectivity”.

Individual-centred statistical techniques require models in which each individual is characterized separately and from which, given adequate data, the individual parameters can be estimated. It is further essential that comparisons between individuals become independent of which particular instruments – tests, or items or other stimuli – within the class

considered have been used. Symmetrically, it ought to be possible to compare stimuli belonging to the same class–measuring the same thing – independent of which individuals within the class were instrumental for the comparison (Rasch, 1960, page vii).

Specific Objectivity is a property taken for granted in the field of physical measurement.

3) The two tests must be equally reliable or perfectly parallel (Lord, 1980). In practice this condition is rarely, if ever, met. A less rigorous definition has been used in connection with constructing statistically equivalent tests.

Non-parallel tests X and Y (that is, tests measuring the same non unidimensional ability but differing in difficulty or reliability) can be considered to be equated if any two examinees of equal true ability, one taking test X and the other taking test Y, would be expected to obtain the same score when performance on test X and test Y are expressed on a common score scale (Kolen, 1981, page 1).

Kolen (1981) referred to the above as the definition of equating for non-parallel tests. Whitely and Dawis (1974) refer to this definition as the equating of "tau-equivalent" measures, where "tau" refers to the symbol "τ" which stands for an ideal true score of a person.

Forms of Equating

There are two general forms of equating, commonly referred to as ***horizontal*** and ***vertical*** equating. Equating test forms that are designed to measure the same attribute at the same difficulty level for the same population is referred to as horizontal equating. In this form of equating, different forms of the test would normally be designed to have comparable item content and similar distributions of item statistics (Slinde & Linn, 1977). Vertical equating, on the other hand, refers to the process of converting to a single scale, scores on forms of a test designed for populations at different educational levels. In contrast to horizontal equating, forms to be vertically equated differ inten-tionally in the difficulty of the items for a single population of examinees and in their content specifications as well (Slinde & Linn, 1977). Thus, the main purpose of vertical equating is to produce a single scale that enables the comparison of ability estimates at different points in time and also to provide information that can be used for comparison of different year cohorts of the same points in time (Hembelton & Swaminathan, 1985; Weiss & Yoes, 1991).

Methods of Equating

Research into theories for the equating of tests, particularly those with items that are dichotomously scored, has been going on for more than 50 years. Test equating methods, according to the testing theory on which they are based, can generally be classified as:

Classical Test Theory Equating: The traditional methods of equating tests revolve around matching the shapes of the distribution of scores. In the case of the linear equating method, the assumption is that the only difference between two tests to be equated is a difference in origin and unit. The linear equating method adjusts for these differences by setting the mean (origin) and standard deviation

(unit) of the same groups of students on the relevant tests to the same means and standard deviation. This type of equating underpins most statistical procedures that are used to moderate school assessments before they are combined with examination scores to produce Tertiary Entrance Scores.

Equipercentile equating method assumes that in general, scores on different tests cannot be equated by adjusting the origin and unit size only. The method requires the cumulative frequency distributions for each test, and assigns the same scaled score to the scores on Test X and Test Y if their percentile ranks are the same. That is, the equivalent scores are scores on Test X and Test Y that have the same percentile rank. This method is generally used in Australian states to adjust for differences among subjects. Once it has been carried out, the scaled scores from the different subjects are added the resulting score is expressed as the Tertiary Entrance Score.

Both of these equating methods assume that the students that have done the two tests are the same students or at least they are randomly equivalent groups. If this is not the case, more advanced equating methods, with additional assumptions must be used. (See Angoff, 1971; Braun and Holland, 1982; Dorans, 1990; Gulliksen, 1950; Marco, Petersen and Stewart, 1983).

Recently, a number of researchers have drawn attention to shortcomings of traditional procedures such as equipercentile and linear methods of equating, particularly for vertical equating tasks, and have suggested that among current methods of equating tests, only those based on item characteristic curve theory (i.e., latent trait models) are appropriate for the tasks of vertical equating (Divgi, 1981; Guskey, 1981; Holms, 1982; Lord, 1975; Loyed &Hoover, 1980; Reckase, 1981; Skaggs & Robert, 1988; Slind & Linn, 1977, 1978, 1979; Smith & Kramer, 1992; Wright, 1977; Wright & Dorans, 1993).

Item Response Theory Equating: The development of Rasch models arose from an equating problem at the level of tests. Reading tests, administered to the same pupils at different stages, to measure the improvement in reading ability (Rasch, 1960/1980) had to be equated. The important characteristic of these unidimensional models for measurement was that they had one parameter for a student, *the ability*, and one parameter for the test, *its difficulty*. Moreover, no assumptions were needed regarding the distribution of student abilities or test difficulties. Thus the student and test parameter, together with the form of the model, were considered to determine the probability of an error in reading each word.

It was from the solution to this problem at the level of the test that Rasch proceeded to the model for dichotomous items, which he later generalized to items with more than two categories. Rasch model for dichotomous items, that is also referred to as Simple Logistic Model (SLM), can be expressed as follows (Andrich, 1988):

$$P\{xni = 1\} = (Bn / Di) / Gni$$

Where

P{xni = 1} is the probability that person n will answer item i correctly

Bn is the location of person on the variable;

Di is the location of item on the variable;

Gni is equal *to* ***1+(Bn / Di)*** and is a normalizing factor that ensures that ***P{xni =1}+P{xni =0} =1.0***

This equation can also be expressed as the logarithmic metric as shown below:

$$P\{x_{ni};\beta_n,\delta_i\} = \exp[x_{ni}(\beta_n - \delta_i)]/\gamma_{ni}$$

Where

βn is the parameter describing the location of person n on the variable;

δi is the parameter describing the location of item i on the variable; and

γni is equal to ***1+exp [βn - δi]*** and is the normalizing factor.

Four data-related conditions have been associated with the requirements that underlie this model:

1. *Local Independence:* the probability that a person responses correctly to a particular item should be independent of the responses that have been made to previous items;
2. *Equality of item discrimination*: the set to which the model is being applied must share a common level of discrimination;
3. *No guessing:* performance on the item set should not be influence by guessing; and
4. *Unidimensionality:* the item set must measure only one trait or ability, (Rentz & Bashaw, 1975). In addition, the Rasch model, as well as other latent trait models, requires the tests to be unspeeded; otherwise, the probability of correctly answering the last items depends not only on the probability of success due to attempting the item but also on the probability of attempting the item (Slinde& Linn, 1979).

One of the advantages of using the method developed by Rasch (1960/1980, 1968, 1977), is that it provides an explicit framework for evaluating the validity of equating any two tests.

When items in different tests have been

- constructed to measure the same property; and,
- shown to fit the requirements of the Rasch model, then they can be transformed onto a single common scale.

Once the items are on a common scale, they share a common calibration. The measures that result from scores on any tests that are drawn from the scale, are automatically equated and no further collection or analysis of data is needed.

Procedures for Equating Using IRT

Traditional methods of test equating with the Rasch model have used common persons or comm.-on items to place items from different tests onto a common scale. There are a number of different estimation procedures for IRT Equating:

1. Concurrent Equating: This procedure uses the overlap between subsets of data to simultaneously estimate item parameters for the Rasch model. This means that the item parameters and the person measures are on the same scale. There is no need to conduct any subsequent transformations and the estimates and measures are directly comparable.

The concurrent equating approach uses the

missing data feature of the recent Rasch programs to determine the locations of the items from the data matrices.

2. Relative Anchoring: This procedure involves calibrating the two tests separately. The parameters in the situation are not necessarily invariant, because in each of the calibrations the item difficulties must be arbitrarily assigned to ensure that the sum is equal to zero (that is, the origin of each of the scales is arbitrary). To ensure that the two tests are on the same scale therefore, differences caused by the arbitrary assignments must be adjusted. The adjustment must be estimated on the basis of having elements in common between the two tests. In the case of common person equating the common elements are the students. In the case of common item equating the common elements are the items that are in both forms of the test. The effect of the difference in local origins is removed by calculating the difference in difficulties between the common items (or persons) from the two tests. A weighted or unweighted average of the differences is used as the link or translation constant necessary to place the items on the same scale.

Equating designs

There are a number of different design or data collection procedures available for equating test scores. The designs have been referred to in a number of different ways in the literature. For the purposes of this paper, however, the names will be referenced to the type of group used in the design. There are three commonly used equating designs:

1. Single group design: In this design two or more forms of a test are administered to the same group of students. Since only one group of students is involved in the test equating process, so the between groups differences are minimised and, consequently, the measurement error is relatively small. This design, however, requires that students answer to a huge amount of items, thus fatigue is the first source of error. Practice effect or test wiseness is the second error source, especially when tests to be equated have the same format. Counter balancing the order of test administration is one possible procedure to eliminate these two sources of error (Kolen, 1988; Kolen and Brennan, 1995).

2. Equivalent group design: In this design two or more forms to be equated are administered to two or more equivalent groups of students. This design is illustrated in Figure 2. As Figure 1 illustrates Group 1 has completed Form A and Group 2 has taken Form B. As assignment of students into Group A and Group B has been randomly this design can be also called the random group design. This design has not got the single group design's disadvantages, namely, fatigue and practice effects. In addition, this design is not time consuming because students take only one test. There is, however, no guaranty that two groups be exactly the same in their ability distributions to eliminate this source of error increasing of sample size is required (Kolen, 1988; Kolen and Brennan, 1995).

3. Anchor-test design: In this design, that is also referred to as the common-item nonequivalent group design (Kolen, 1988), two or more forms to be equated are administered to two or more different groups of students. In contrast with the

equivalent group design, groups can have different ability distribution. The other feature of this design is that a set of common items is used to adjust the difference between the test forms. Because groups may have a different ability distribution, this design has widely used in longitudinal studies to measure growth spurt. This design is also extremely useful in developing item banks.

This Study

There have been numerous studies that focus upon effectiveness of equating using IRT. Most of these have focused upon evaluating the utility of various equating designs and procedures where the data fit the Rasch mode (e.g. Slinde & Linn, 1977, 1978, 1979; Gustafsson, 1979; Loyd & Hoover, 1980; Holmes, 1982; Schratz, 1984; Shen, 1993). However, in practice, data rarely fit the model closely. This study, therefore, tends to take these earlier studies a step further by examining what happens to the results that evolve from using different equating procedures, as the fit to the Rasch model gets progressively worse. More particularly this study is designed to compare the results generated from two equating procedures (concurrent and relative anchoring) applied to the same data sets as fit to the Rasch model progressively deteriorates.

Methodology

For the purposes of this study, data were generated to fit the rasch Model using Simulation Program for Rasch Data or RUMMSims (Andrich, Luo & Sheridan, 1997). These data were then adjusted to generate datasets that fit the Rasch Model less well. The following method has been used to produce and compare the equating results from the two equating procedures.

Data fit the model: The following procedure is used to generate a data set that has relatively good fit to the Rasch Model; split the data set into two test form data sets with items in common; equate the two test forms using the concurrent and relative equating procedures; and compare the results obtained from the two equating procedures.

1. Use the ***RUMMSims*** program to generate a data set.
2. Divide the data set into two separate Tests (Test A and Test B) with 30 items in each. Tests A and B are linked by 10 common items (item 21 to 30 in Test A; items 31 to 40 in Test B).
3. Use the *RUMM* (Rasch Unidimensional Measurement Model) program to estimate item and ability locations for Test A and Test B separately.
4. Use the RUMM program and employ *relative* and ***concurrent*** equating to place the item difficulty estimates from one test on the same scale as the other.
5. ***Relative equating*** involves the calculation of the mean of the differences of the difficulties of common items in the two tests being equated. The results are then used to place one scale on the other scale.
6. ***Concurrent equating*** involves pooling the data generated by the two tests into one combined data set. This process leads to the simultaneous calibration of items of both tests onto the one scale.

7. Compare the equating results obtained from the different equating procedures by plotting them on the same graph with the 95% confidence intervals.

These plots are used to evaluate the invariance of item difficulty or person measures and hence the quality of the items. 95% confidence intervals make it easy to see how satisfactorily the item points in the plots follow the expected identity lines.

If the data fit the measurement model, then it is expected that the independent estimates of the difficulties of the items that emerge from different equating procedures will be statistically equivalent. The small differences between two scales show that different equating procedures can produce results equivalent to a combined calibration of both Test A and Test B.

Data do not fit the model: The following procedure shows the different steps that are carried out in preparing and comparing equating data when those data are not in accordance with the model.

1. Use the ***RUMMSims*** to generate a data set.
2. Change the content of data sets generated by the Rasch model to produce data sets that do not fit the Model.
3. Total – Item Chi-square, generated by RUMM program as Item-trait Interaction, is used to pass judgment about the degree of fit-the-model. Three groups have been selected since Chi-Square is affected by the number of groups used to conduct a test-of-fit, three groups have been selected to conduct a test-of-fit. Degree of fit-the-model is divided into five categories as follow: **1)** Very Good Fit (VGF), **2)** Good Fit (GF), **3)** Average Fit (AF), **4)** Poor Fit (PF), and **5)** Very Poor Fit (VPF).
4. Divide the data set into two separated Tests (Test A and test B) with 30 items in each. Tests A and B are linked by 10 common items (item 21 to 30 in Test A; items 31 to 40 in Test B).
5. Repeat step 3 to 7 above.

Any large differences between two scales show that different procedures of equating can produce different results in comparison with a combined calibration of both Test A and Test B.

Results

In this study the anchor-test design (e.g. common items nonequivalent design) is employed to equate two tests of different difficulty that have been given to two groups of different ability level. This section presents the results obtained from applying different equating procedures, as the fit to the Rasch model gets progressively worse.

As mentioned earlier, the present study compares the equating results generated from different equating procedures by plotting them on the same graph with their 95% confidence intervals. These plots are used to evaluate the invariance of item difficulty or person measures obtained from different equating procedures under different model-data fit. If data fit the Rasch model relatively well, then it is expected that the independent estimates of item fit that emerge from the two different equating procedures should be the same or at least relatively equivalent.

Any small differences that do emerge between the two estimates indicate that different equating

procedures produce equivalent results and there is no problem with alternating and using different equating procedures. However, if there are large differences between these two estimates produced from the different equating procedures then it means that the equating procedures do give different results and that using different equating procedures on different occasions can have significant implications for the problem. For example, if a student is above a selection cut-score when two tests are equated using relative anchoring, then there are serous problems with interpreting the results.

Figure 1 shows the scatter plot of the person measures produced after equating two tests using concurrent equating and relative anchoring procedure when the data in both tests fit the Rasch model.

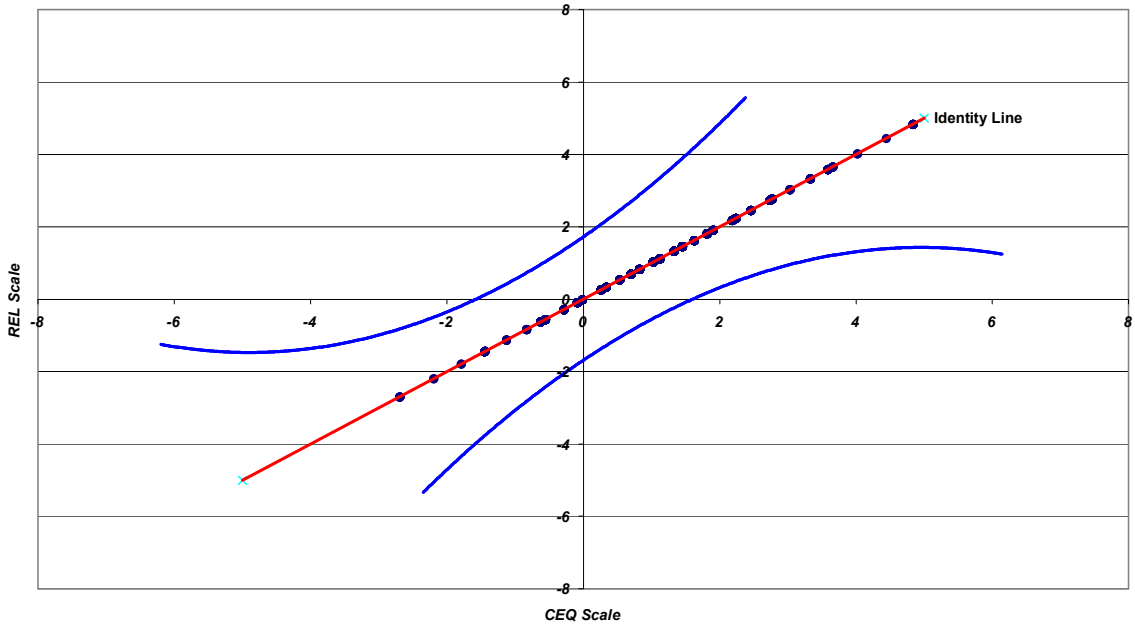

Figure 1: Comparison of Concurrent Equating with Relative Anchoring Procedure based on Person locations. *(n=500; i=20; ci=10)*

As can be seen, all person measures lie within the 95% confidence limits. It means that both concurent equating and absolute anchoring have generated the same equating results.

Another way to examine the educational significance of the results is to use a mis-classification table as shown in Table 1.

Table 1: Comparison of Concurrent Equating with Relative Anchoring Procedure Results in terms of Cut-Score=31

Relative Anchoring	Concurrent Equating	
	Above Cut-Score	Below Cut-Score
Above Cut-Score	65	0
	6.5%	*0.0%*
Below Cut-Score	0	935
	0.0%	*93.5%*

This Table shows that the number of students above and below the cut-score is exactly the same. It means that there is no difference between results generated from concurrent equating and relative anchoring procedures.

Figure 2 illustrates the same scatter plot when the model-data fit in Test A is very good but the model-data fit in Test B is very poor.

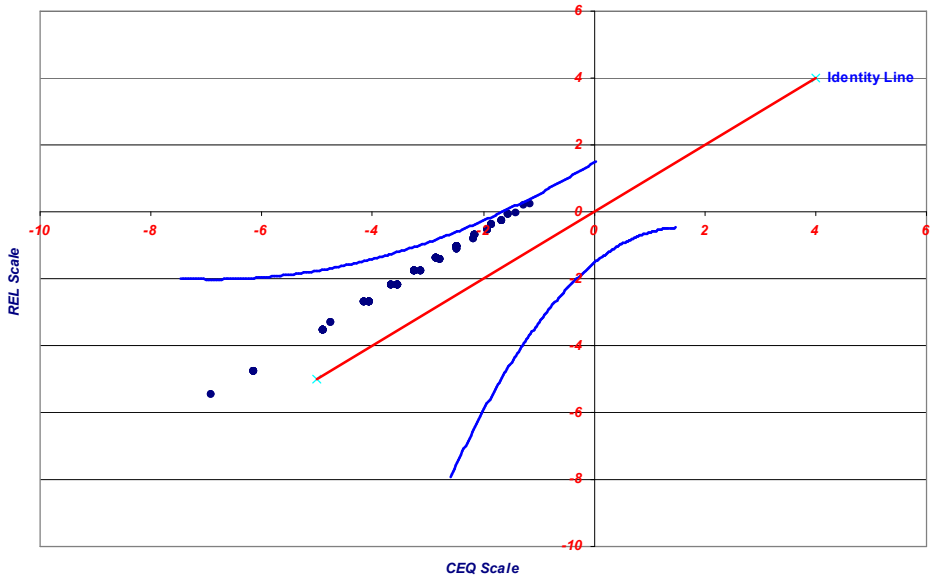

Figure 2: *Comparison of Concurrent Equating with Relative Anchoring Procedure based on Person locations.*

(n=500; i=20; ci=10)

As this Figure reveals, the person measures emerged from the different equating procedures provided different results. Based on this scatter plot, it seems clear that two equating procedures, concurrent equating and absolute anchoring, have generated slightly different equating results as the fit to the model has degenerated.

Table 2: Comparison of Concurrent Equating with Relative Anchoring Procedure Results in terms of Cut-Score=0

Relative Anchoring	Concurrent Equating	
	Above Cut-Score	Below Cut-Score
Above Cut-Score	0	2
	0.0%	*1.0%*
Below Cut-Score	0	198
	0.0%	*99.0%*

The equating results obtained by these two procedures relative to an arbitrary cut-score of 0, are presented in Table 2. As this Table indicates 2 students would have their classification changed according to the use of the equating procedure. That is two students who would be classified as being below the cut score if the data were equated using concurrent equating would in fact be above the cut score if the tests were equated using relative anchoring.

Figure 3 shows the results when the model-data fit in Test A is very poor and the model-data fit in Test B is good.

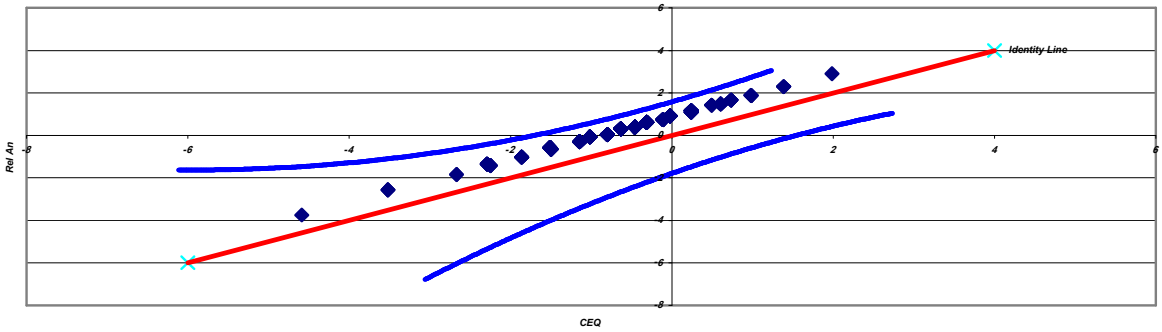

Figure 3: *Comparison of Concurrent Equating with Relative Anchoring Procedure based on Person locations.*

(n=100; i=20; ci=10)

As this Figure reveals, the person measures emerged from the different equating procedures are different. It means that two equating procedures, concurrent equating and absolute anchoring, have not generated the same equating results.

The equating results obtained by these two procedures, relative to an arbitrary cut score of 1.5, are presented in Table 3.

Table 3: Comparison of Concurrent Equating with Relative Anchoring Procedure Results in terms of Cut-Score=1.5

Relative Anchoring	Concurrent Equating	
	Above Cut-Score	Below Cut-Score
Above Cut-Score	1	13
	0.5%	*6.5%*
Below Cut-Score	0	186
	0.0%	*93%*

As this Table shows 13 students would have their classification changed according to the use of the equating procedure. In other words, 13 students who are above the cut score, when applying relative anchoring, would be classified as being below the cut score if two tests were equated using concurrent equating. In this situation, the students that their marks have been equated by applying concurrent equating procedure would be at a disadvantage relative to students that relative anchoring procedure has been applied to equate their marks.

The scatter plot of the person measures produced after equating two tests using concurrent equating and the relative anchoring procedures when the model-data fit in both tests is very poor is illustrated in Figure 4.

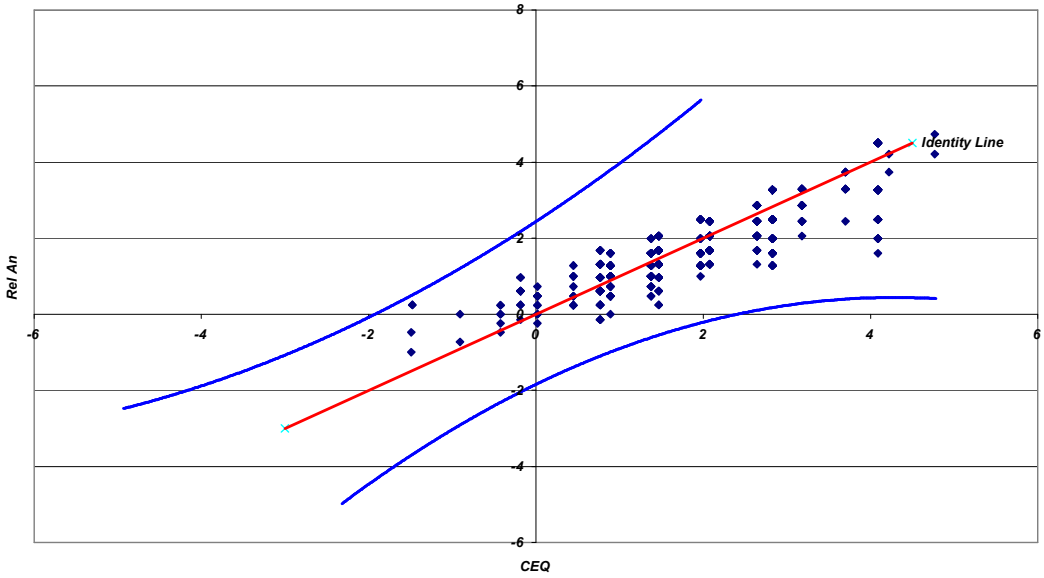

Figure 4: *Comparison of Concurrent Equating with Relative Anchoring Procedure based on Person locations.*

(n=250; i=20; ci=10)

As can be seen, the person measures emerged from the different equating procedures generated significant variations. It means that two equating procedures, concurrent equating and absolute anchoring, have generated very different equating results as the fit to the model in both tests has degenerated.

Table 4: Comparison of Concurrent Equating with Relative Anchoring Procedure Results in terms of Cut-Score=3

	Concurrent Equating	
Relative Anchoring	**Above Cut-Score**	**Below Cut-Score**
Above Cut-Score	44	17
	8.8%	*3.4%*
Below Cut-Score	33	423
	6.6%	*84.6%*

The equating results emerged from concurrent equating and relative anchoring procedures, relative to an arbitrary cut score of 3, are presented in Table 4. As this Table shows 50 students would have different classification according to the use of the equating procedure. That is 17 students who are above the cut score, when applying relative anchoring, would be classified as being below the cut score if two tests were equated using concurrent equating. Similarly, 33 students would who would be classified as being below the cut score if the data were equated using relative anchoring would be above the cut score if the tests were equated using concurrent equating procedure.

Figure 5 illustrates the scatter plot of the person locations produced after equating two tests using concurrent equating and the relative anchoring procedures when the model-data fit in Test A is poor and the model-data fit in Test B is very good.

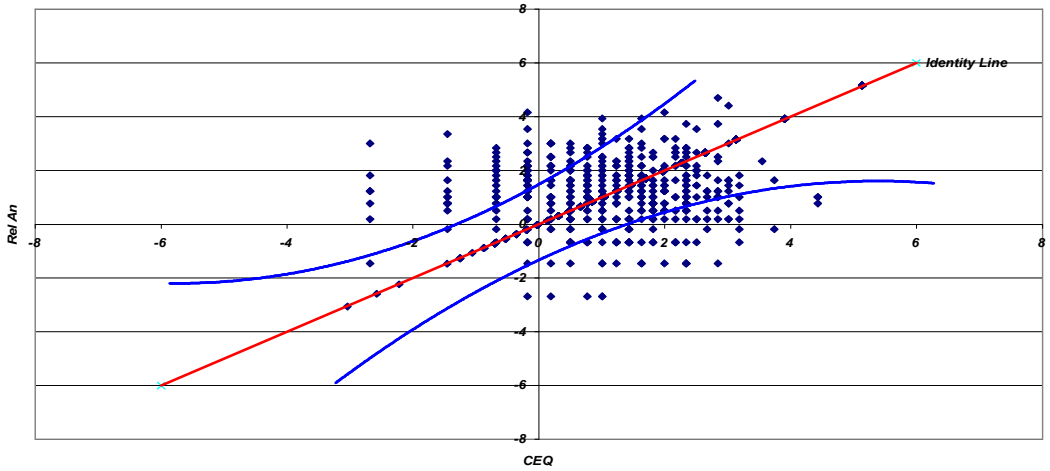

Figure 5: *Comparison of Concurrent Equating with Relative Anchoring Procedure based on Person locations.*

(n=500; i=30; ci=10)

As this Figure reveals, the majority of person locations lie outside the 95% confidence limits. It indicates concurrent equating and relative anchoring, have generated very different equating results.

The equating results obtained by these two procedures, relative to an arbitrary cut score of 3, is shown in Table 5.

Table 5: Comparison of Concurrent Equating with Relative Anchoring Procedure Results in terms of Cut-Score=3

	Concurrent Equating	
Relative Anchoring	**Above Cut-Score**	**Below Cut-Score**
Above Cut-Score	46 *4.6%*	28 *2.8%*
Below Cut-Score	21 *2.1%*	905 *90.5%*

As this Table shows 49 students would have

their classification changed only because of the using of equating procedure. In this case, 28 students who are above the cut score, when using relative anchoring, would be classified as being below the cut score if two tests were equated by using concurrent equating. Similarly, 21 students would who would be classified as being below the cut score if the data were equated using relative anchoring would be above the cut score if the concurrent equating procedure were used to equate to tests.

Discussion

The purpose of present study is to investigate the relationship between different procedures of test equating and fit of the data to the Rasch model. The findings indicate that when data fit the Rasch model there is no significant difference between equating results.

From a theoretical perspective, these results are to be expected. If two tests fit the model relatively well, then the resulting student scores should be independent of the items that the students attempt. If the tests do not fit the model as well, then the concurrent equating procedure will take the total set of items in the two tests and produce a variable that is an amalgam of the two tests. The variable is not the same as either of the two tests. The relative anchoring on the other hand produces a score that is a direct result of adding a translation constant to one of the tests. The variables have not been merged in any "real" way to produce a new variable. In fact the equating results become less conceptually comparable as the fit of the data to the model diminishes.

As fit to the model gets progressively worse the results from using different equating procedures are less comparable and the consequences from a selection point of view become less acceptable.

When this result is translated to a situation in Australia, for example, where different states use different equating procedures to generate a single state scale and then these scales are amalgamated onto a national scale for the purposes of comparing student performance to predetermined standards or benchmarks, it raises significant equity issues. Would the same students or number of students designated, as being below benchmark, remain the same? The answer to this is yes, if the state data fits the model relatively well and no, if the state data are less in accord with the Rasch model.

References:

Andrich, D. (1988). *Rasch Models for Measurement*. Newbury Park, California: Sage.

Andrich, D., Luo, G., & Sheridan, B. (1997). *RUMMSims: Simulation Program for Rasch Data*, RUMM Laboratory Pty Ltd.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508-600). Washington DC: American Council on Education.

Beguin, A. A. (2000). Robustness of equating high-stakes tests. Druk: FEBODRUK B. V., Enschede.

Divgi, D. R. (1981). Model-free evaluation of equating and scaling. *Applied Psychological Measurement*, 5, 203-208.

Guskey, T. R. (1981). Comparison of a Rasch model scale and the grade-equivalent scale for vertical equating of test scores. *Applied Psychological Measurement*,

5, 187-201.

Gustafsson, J. E. (1979). The Rasch model in vertical equating of tests: A critique of Slinde and Linn. *Journal of Educational Measurement*, 16(3), 153-158.

Hembelton, R. K., & Swaminathan, H. (1985). Item *Response Theory: Principles and Application*. Kluwer Academic Publishers, Boston, MA.

Holmes, S. E. (1982). Unidimensionality and vertical equating with the Rasch model. *Journal of Educational Measurement*, 19, 139-147.

Keeves, J. P. (1992). Scaling achievement test scores. *In The IEA Technical Handbook*. IEA, The Hugue.

Kolen, M. J. (1988). An NCME instructional module on traditional equating methodology. *Educational Measurement: Issues and Practice*, 7, 29-36

Kolen, M. J., & Brennan, R. L. (1995). *Test Equating: Methods and practices*. New York: Springer.

Lord, F. M. (1975). *A survey of equating methods based on item characteristic curve theory* (ETS RB 75-13). Princeton NJ: Educational Testing Service.

Lord, F.M. (1977). Practical application of item characteristic curve theory. *Journal of Applied Measurement*, 14(2), 117-139.

Lord, F.M. (1980). Applications of Item Response Theory to Practical Testing Problem. Hillsdale, New Jersey: Erlbaum.

Lord, F.M., & Stocking, M. L. (1988). Item Response Theory. In J. P. Keeves (Ed.), *Educational research, methodology and measurement: An international handbook*. Pergamon Press, Oxford.

Loyed, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, 17, 179-193.

Peterson, N.S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming and equating.In R.L. Linn (Ed.), *Educational measurement* (3rd. Eition), Macmillan, New York.

Rasch, G. (1960/1980). *Probabilistic model for some intelligence and attainment tests*. Danish Institute for Eucational Research, Copenhagen.

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, 19, 49-57.

Rasch, G. (1968). *A mathematical theory of objectivity and its consequences for model construction*. In "Report from the European Meeting on Statistics, Econometrics and Management Sciences", Amsterdam.

Reckase, M.D. (1981). *To use or not to use (the one or three-parameter logistic model) –That is the question*. Paper presented at the Annual Meeting of the American Educational Research Association (65th. Los Angeles, CA, April 13-17, 1981). Office of Naval Research, Arlington, Va. Personnel and Training Research Program Office, Missouri University, Columbia.

Rentz, R. R., & Bashaw, W. L. (1975). *Equating reading tests with Rasch model (Vols. 1 & 2)*. Athens: University of Georgia, Educational Research Laboratory. (ERIC Document Reproduction Service Nos. ED 127 330-ED 127 331).

Schratz, M. K. (1984). *Vertical equating: An empirical study of consistency of Thurstone and Rasch model approaches*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, La, April 24-26, 1984.

Shen, L. (1993). *Constructing a measure for longitudinal medical achievement studies by the Rasch model one-step equating*. Paper presented at the Annual Meeting of the American Educational Research association, Atlanta, GA, April, 1993.

Skaggs, G., & Robert, W. (1988). Effect of examinee ability on test equating invariance. *Applied Psychological Measurement*, 12(1), 69-82.

Slinde, J. A., & Linn, R. L. (1977). Vertivally equated tests: Fact or phantom?. *Journal of Educational Measurement*, 14, 23-32.

Slinde, J. A., & Linn, R. L. (1978). An exploration of the adequacy of the Rasch model for problem of vertical equating. *Journal of Educational Measurement*, 15, 23-35.

Slinde, J. A., & Linn, R. L. (1979). The Rasch model, objective measurement, equating and robustness. *Applied Psychological Measurement*, 3, 437-452.

Smith, R.M., & Kramer, G. A. (1992). A comparison of two methods of test equating in the Rasch model. *Educational and Psychological Measurement*, 52(4), 835-846.

Weiss, D. J., & Yoes, M. E. (1991). Item response theory. In R.K. Hambletone and J. N. Zaal (Eds.), *Advances in education and psychological testing*. Kluwer Academic Publishers, Boston, MA.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14, 97-116.

Wright, B. D. (1995). 3LP or Rasch? *Rasch Measurement Transactions*, 9(1), 408-409.

Wright, N. K., & Dorans, N. J. (1993). Using *the selection variable for matching or equating.* Educational Testing Service, Princeton, N. J.