INDIAN CENTRE FOR ASSESSMENT, EVALUATION AND RESEARCH (CAER)

Some Possible Options for Aggregating Subject Examinations at the Level of Scores for Entry into Indian Institutes of Technology

---

Jim Tognolini and Jon Twing

9/28/2012

This paper provides some options for scaling the examination scores across Boards and subjects so that students are neither advantaged nor disadvantaged by the group of students that are being compared within their Board. The paper will also provide some data to show some of the effects that can occur to the rank order if the scaling is done.

Some Possible Options for Aggregating Subject Examinations at the Level of Scores for Entry into Indian Institutes of Technology

1.0 Introduction

The Joint Institute of Technology Joint Entrance Examination (JEE) is an annual entrance examination for the 16 Indian Institutes of Technology (IITs). It has been used for entry since 1960[1]. In the early years it was called the Common Entrance Examination (CCE) and was initiated in response to the IIT Act of 1961.

There have been a number of changes and reforms to the test over the years. The most recent changes have just been accepted and will be implemented in 2013. The latest structure will comprise 2 JEE Examinations; JEE (Main) and JEE (Advanced). Candidates wishing to get entry into an Indian Institute of Technology (IIT) will have to sit for the (JEE Main). The students who perform best on this examination (approximately 1.5 lakhs) will then be eligible to sit for the JEE (Advanced) exam.

Successful candidates will be those who are in the top 20 percentile of their board examinations AND have JEE (Advanced) scores high enough in the rank order of applicants to qualify for the number of positions available.

In addition, the JEE-Main (which was until 2012 known as the All India Engineering Entrance Examination (AIEEE) is also to be used for admission to various Central Engineering Institutes other than IITs. The final rank list for entry into these (non-IIT) institutes will be prepared by giving 40% weighting to ("normalised") Grade XII Board examination scores and 60% to scores on the JEE-Main. This paper provides some options for "normalising", "scaling" or "equating" the Board marks so that they can legitimately be compared and contribute, along with the JEE-Main score towards a Tertiary Entrance Score.

This change has enhanced the importance of the examinations conducted by the various Boards across India; because students have to perform well in their Board examinations (be in the top 20 percentile) in order to be eligible for entry.

However, this decision has also introduced an issue regarding the comparability of the Board examinations. Given that there is going to be a single rank order of merit produced for entry into IITs, it is problematic to just take the top 20 percentile of students from the various Boards across the country because there has been no attempt to adjust for the relative differences in the "ability" of various cohorts or for the differences in "difficulty" of various content or subject examinations. In other words, it may be a lot easier to beat 20 percent of the cohort of students in one Board than it is in another, particularly if the subjects are not the same; to just take the top 20% from each Board without adjusting for the "ability" of the comparable cohort would not be considered "fair" by most communities and so there is a need to scale the results to produce a single rank order of merit of examination results across the country and the top 20 percentile of students on this scaled result would have met one of the eligibility criteria for entry into IITs.

This paper provides some options for scaling the examination scores across Boards and subjects so that students are neither advantaged nor disadvantaged by the group of students that are being compared within their Board. The paper will also provide some data to show some of the effects that can occur to the rank order if the scaling is done.

[1] It was originally referred to as the Common Entrance Examination (CEE).

2.0 Background to Scaling

A serious limitation of traditional testing is that the "raw" score that a student obtains on the test (typically obtained by summing the marks the student obtains on the test) can only be interpreted in terms of the particular test that has been used. When a test is constructed to assess performance on a subject, the test questions or the items that are used represent only a sample of the possible set of items that could have been used. The score obtained by the student is dependent on the particular items chosen for that particular examination. If a different set of items (or even a few different items) had been taken, a different score would probably have been obtained and hence a different rank ordering of students on raw score might be realised. To this end the score on the subject is peculiar to the set of items through which it is defined.

In many testing situations it is necessary to compare the scores of students who have taken different forms of a test. One of these situations occurs when the test contains a choice of items. In practice, the tests cannot be expected to be of equal difficulty for students at all ability level. Therefore, a comparison of total scores of the different forms of the test would not be fair to the students who have taken the more difficult set of items.

A second situation and one that is more the focus for this paper occurs when scores obtained from different combinations of subjects need to be compared and included in the calculation of a student's single Tertiary Entrance Score (TES). The scores in this situation are defined in the metric of the subject (and as such are not strictly comparable) and would depend upon the relative difficulty of the subject for the particular cohort of students attempting it. In India and the UK there is an added layer of complexity in that the examinations have been carried out by different Boards and there is no alignment of difficulty, across the Boards, for examinations in the same subject; or, across different subjects.

There are numerous more situations ranging from monitoring the performance of students over time when tests that differ in context and difficulty are used in different years (e.g. Program for International Student Achievement [PISA} and the International Mathematics and Science Surveys [e.g. TIMSS]); tocomparing school assessments prepared by different schools across the country. In this latter case in order that valid comparisons among students from different schools can be made, different teachers' assessments must be placed onto a common scale before they are compared.

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 report to provide a survey of all of the topics associated with test equating. However, it is the aim of this report to provide an overview of some of the more common methods for linking different tests and examinations.

3.0 Equating, Scaling and Linking Tests and Examinations

Equating, scaling and linking are terms used to describe the empirical procedures used in transforming the scores of tests and examinations 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.

Bèguin (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; Petersen, Kolen and Hoover, 1989).

> 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.
>
> (Bèguin, 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 inherent in these definitions. The first is that the tests to be equated must be measures of the same variable. 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 mathematics test from the Central Board of Secondary Education (CBSE) to one from another Indian Board 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 mathematics to a test, which is a measure of artistic aptitude, is worthless.[2]

The second condition implied by equating is that 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:

[2]Angoff's premise might be construed to mean that linking a mathematics problem-solving test with a mathematics number sense assessment would not be allowed because the two tests measure different things. However, both tests measure aspects of a higher order variable (i.e., mathematics skill) such that the linking becomes plausible. Similarly, while it might not make sense to link an art test with an arithmetic test, if these two scores will be used as indicators of a higher order measure (e.g. Tertiary Entrance Scores) the need or requirement is not so clear.

> …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; namely no scientist asks which thermometer is used to measure temperature - it is just assumed the thermometers are calibrated before measures are taken.

A third condition proposed by Lord (1980) is that the two tests must be equally reliable or perfectly parallel. 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.

Test equating procedures are generally classified into 2 categories: horizontal and vertical equating.

Horizontal equating is used to describe the process of equating 2 or more tests that are designed to measure the same property at the same academic level. Vertical equating is the equating procedure that is used to equate tests that measure the same property at different academic levels (Holmes, 1982; Mislevy, 1992).

4.0 Methods of Equating

4.1 Classical Test Theory (TTT) equating

Research into methods of equating tests (particularly those with items that are dichotomously scored) has been going on for more than 50 years. Many different methods for test equating have been proposed but until the advent of latent trait methods, the linear and equipercentile equating methods had been the most commonly used and most research, at least until the last decade, has focused extensively on them.

The traditional methods of equating tests revolve around matching the shapes of the distribution of scores. In the case of the ***linear scaling method***, the assumption is that the only difference between two tests to be equated is a difference in origin and unit. The ***linear scaling method*** adjusts for differences by setting the mean (origin) and standard deviation (unit) of the same groups of students to the same means and standard deviations on the equating variables. 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.

The ***equipercentile scaling 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 TEST). While linear linking establishes equivalence between means and standard deviations, equipercentile scaling extends this linking such that all four central moments (mean, standard deviation, skew and kurtosis) are equivalent.

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 Gulliksen, 1950; Braun and Holland, 1982; Angoff, 1971; Dorans, 1990; and, Marco, Petersen and Stewart, 1983).

4.2 Rasch 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.

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 (subjects taken by students) 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.

5.0 Scaling Examination Results from Different Indian Examination Boards

The basic problem confronted in this paper is the following: Changes to the requirements for IIT entry from 2013 onwards sees a situation whereby candidates who successfully obtain a place at a tertiary institution including some IITs will be those who are in the top 20 percentile of their board examinations and have JEE (Advanced) scores high enough in the rank order of applicants to qualify for the number of positions available.

This change has enhanced the importance of the examinations conducted by the various Boards across India; because students have to perform well in their Board examinations (be in the top 20 percentile) in order to be eligible for entry.

The suggested change highlights an equity issue that currently exists in the system but has never been a problem because there has never been a need to identify the top 20 percent of candidates **across the country**.

It would be problematic (unfair) to just take the top 20 percentile of students from the various Boards across the country because there has been no attempt to adjust for the relative differences in the “ability” of the cohorts that comprise the Boards candidatures; nor has there been any consideration regarding the differences in difficulty of various subjects taken. It would be easier to get into the top 20 percent of a Board candidature that is less able than that from another Board. Given the high stakes nature of the selection process it would seem to be necessary to adjust for the differences in the “ability” of the candidatures before comparing the results across Boards.

In addition, since different students can take different subjects to produce their final scores and these subjects can vary in difficulty then it seems to be necessary, in the interests of fairness, to take account of the relative difficulty of subjects before adding the subject scores to get a final score. It would be unfair for someone who sat the “easiest” subjects to gain entry ahead of someone who had chosen the most “difficult” subjects.

Ultimately the aim should be to ensure that students are not advantaged nor disadvantaged by their subject choice; nor advantaged or disadvantaged by the Board that set the examinations.

It is with this basic principle of equity that the following optional scaling procedures are considered.

Option 1

One option is to just identify the score obtained by the top 20 percent of the students on the examinations conducted by the particular Board and the pool of eligible students would be those in the top 20 percentile irrespective of the Board. This option was explored in the Hindustan Times which stated the following:

> *"In a quirky scenario, a student of the West Bengal class 12 board will need just 58% to be eligible to take the IIT-Joint Entrance Exam (JEE) next year while an aspirant from the Tamil Nadu board will have to score nearly 78% to make the cut.*
>
> *Preliminary data of seven boards across the country shows that the percentage required to be in the top 20 percentile — a necessary condition to be eligible for IIT-JEE next year — will vary for different boards".*
>
> (Hindustan Times, 16 July 2012)

The major problem with this method of identifying the top 20 percentile of students is that it assumes that the populations of students from the various Boards areof equal ability. This assumption is difficult to sustain and is problematic to say the least. It is not fair to students who are competing with more able candidatures and could lead to students moving from one jurisdiction (Board) to another to maximise their chances of meeting the basic criterion for entry. An additional problem is related to the fact that within the jurisdiction of an Examination Board the student Final Score (on which their percentile is based) is generally obtained by adding the scores on 5 subjects and the subjects are not all the same (in other words there is choice in the subjects candidates can choose); and, there has been no attempt to account for the relative difficulty of the different subjects.

Option 2

A second option which could address the issue (different entry scores across Boards) identified in the Hindustan Times would involve first "normalising" the distributions from the various examination Boards so that they all had a mean of zero and a standard deviation of 1; and then giving them a "common" mean and standard deviation so that the scores for the students on the 20th percentile would be exactly the same across the Boards. This would mask the anomaly identified by the journalist, but would not overcome the problem associated with the comparability of the cohorts.

Option 3

The next group of options does address the measurement issue prevalent in the first two options in that they use a common test to scale for differences across subjects and across Boards. The **All India Engineering Entrance Examination (AIEEE)** is the test that is done by significant numbers of students from National and State Examination Boards throughout India (It will be replaced by the JEE (Main) in 2013). It can be used to take account of the general ability of the cohort of students taking each subject by scaling the distribution of the students in the various subjects (irrespective of Board) to the distribution of the scores of the same group of students on the AIEEE. A national test of this nature enables the Board differences and the subject differences to be taken care of concurrently.

Option 3, therefore, involves using a **linear scaling method**to adjust the mean (origin) and standard deviation (unit) of the students ***in each subject***to the same means and

standard deviations of those same students on the scaling test (the AIEEE). Using this transformation, every score on the subject examination can be converted to an equivalent score on the AIEEE; and as a consequence, be directly compared. The following equation can be used for the transformation.

$$Y_{psb} = \left\{ \left[\frac{(X_{psb} - X_{sb})}{\sigma_{sb}} \right] x\, \sigma_{A_{sb}} \right\} + A_{sb}$$

where Y_{psb} is the scaled subject score 's' for person 'p' within board 'b';
X_{psb} is the examination score in subject 's' for person 'p' within board 'b';
X_{sb} is the mean examination score for all persons attempting subject 's' within board 'b';
σ_{sb} is the standard deviation of the examination scores for all persons attempting subject 's' within board 'b';
$\sigma_{A_{sb}}$ is the standard deviation of the AIEEE 'A' scores of the all of the persons taking subject 's' within board 'b'; and,
A_{sb} is the AIEEE 'A' mean of all the students taking subject 's' within board 'b'.

One of the problems with this method of equating is that the scaled scores can have values below the minimum score of the scaling test and above the maximum score. This can be adjusted by fixing the minimum scaled score to zero and the maximum scaled score to 100. This scaling method adjusts for the mean and spread of the distribution but primarily retains the features of the examination scores. If on the other hand it is desirable for the scaled scores to conform closely to the distribution of the scores on the scaling test (AIEEE) then it is essential to equate the percentile distributions.

With the particular problem outlined above it is important that the distribution of scores on the examinations better match the distribution of scores on the AEIII so an equipercentile scaling procedure would be more appropriate. The following option shows an Equipercentile Scaling Method for solving the problem.

Option 4

Option 4 is referred to as the ***equipercentile scaling method***. It assumes that in general, scores on different tests cannot be equated using just the mean and standard deviation only. It requires the cumulative frequency distributions for each test, and assigns, for equivalent percentile ranks on the two, the same scores on the subject examination as on the AIEEE. Once this scaling has been carried out, the scaled scores for the different subject examinations are comparable.

In order to illustrate equipercentile scaling, consider the following steps in which the scores from mathematics from board are scaled onto the AIEEE distribution.

STEP 1: Determine the score 's' in the AIEEE (ALL) distribution that corresponds to, for example, to a percentile rank of 50.

STEP 2: Convert the score's' to a percentile rank 'p' in the AIEEE (subject) distribution.

STEP 3: Convert the percentile rank 'p' to a score 'e' in the subject distribution.

STEP 4: The ordered pair (e, s) is used as one of the points in the equating process.

STEP 5: A similar strategy is used to equate all of the subject scores to the AIEEE scale.

Once this process has been done for all subjects across all boards, scores in any two subjects are considered equivalent if they correspond to the same score on the AIEEE. Furthermore, it is considered that they can be added and the resulting total score used to summarise the overall performance of candidates on a common scale.

After scaling, it is considered appropriate to answer the question, "What is the score in mathematics in the Board x examination that corresponds to a score of 75 in another subject from another Board?" It is also considered that the aggregate that results from the scaled scores can be compared directly and it is possible to ascertain the top 20 percentile of candidates across India irrespective of which examinations they used to generate the aggregate and which examination Board administered the examination.

Illustrative Example

The data set that is used to illustrate equipercentile equating started with the total pool of students (in excess of 185,000) and all subjects.

The first step of the analysis was to remove subjects with insufficient numbers of students. It was arbitrarily decided that any subjects with less than 100 candidates would be removed from further analysis for the purposes of this example. This resulted in 39 subjects being available for the analysis. It was decided (again arbitrarily) to focus on subjects). As such, Table 1 provides the names and a brief description of the remaining 35 subjects to be analysed.

TABLE 1

Subjects used in Illustrative Example

Variable Name	Description	Note
BENG_TOT	AIEEE TOTAL SCORE	Engineering Entrance Exam
MRK_041	FUNC-ENG	Language
MRK_042	PUNJABI	Language
MRK_043	BENGALI	Language
MRK01	ENGLISH ELECTIVE	Language
MRK02	HINDI ELECTIVE	Language
MRK07	GEOGRAPHY	Course Content
MRK08	ECONOMICS	Course Content
MRK11	MUSIC HIND.VOCAL	Music
MRK12	MUSIC HIND.INS.MEL	Music
MRK14	PSYCHOLOGY	Course Content
MRK16	MATHEMATICS	Course Content
MRK17	PHYSICS	Course Content
MRK18	CHEMISTRY	Course Content
MRK19	BIOLOGY	Course Content
MRK20	BIOTECHNOLOGY	Course Content
MRK21	ENGG. GRAPHICS	Course Content
MRK22	PHYSICAL EDUCATION	Course Content
MRK23	PAINTING	Course Content
MRK26	APP-COMMERCIAL ART	Course Content
MRK33	HOME SCIENCE	Course Content
MRK34	INFORMATICS PRAC.	Course Content
MRK35	ENTREPRENEURSHIP	Course Content
MRK36	MULTIMEDIA & WEB T	Course Content
MRK40	COMPUTER SCIENCE	Course Content
MRK41	FUNCTIONAL ENGLISH	Language
MRK42	PUNJABI	Language
MRK43	BENGALI	Language
MRK46	MARATHI	Language
MRK48	MALAYALAM	Language
MRK51	KANNADA	Language
MRK62	ENGLISH CORE	Language
MRK63	HINDI CORE	Language
MRK65	SANSKRIT CORE	Language
MRK66	TYPOGRAPHY &CA ENG	Course Content

After the calculation and inspection of the descriptive statistics were completed, each subject was statistically linked to the AIEEE total score BENG_TOT using the procedure outlined above.

The linking method used a common type of statistical test form equating commonly known as the “randomly-equivalent groups design” (Kolen& Brennan, 2004). The linking used no smoothing. While the assumption of randomly equivalent groups is not a strong assumption for the current linking given that the linking was performed on subsets of data with common persons, the ability to generalise this linking is strong. In

other words, if larger or different groups of candidates were used, different conversion tables between each subject and the AIEEE total score would be realised.

Using the linking methodology previously described, each subject (see Table 1) was equated to the AIEEE total score (variable BENG_TOT) using the R package "equate" (Albano, 2011). As such, a total of 34 "pair-wise" linkings were completed. Each individual linking resulted in a conversion table, which provided an AIEEE total score equivalent for each individual subject scale score. For example, when the 0-100 scale of MRK01 was equated to the -51 to 345 scale for BENG_TOT two column vectors resulted. In one column was the BENG_TOT total scale score and in the other was the MRK01 scale score equivalent.

In this way a simple conversion of MRK01 (which in this case is the English Elective course for language study) to the BENG_TOT scale (AIEEE total) has been made. Such an equipercentile linking was then performed for all of the remaining subjects such that each subject had a linked AIEEE total score equivalent.

A comparison of a composite score obtained from a simple average of these "equated scores" with the empirically obtained AIEEE shows the differences possible from decisions resulting from the composite score relative to decisions made from the use of the AIEEE score only. Furthermore, a simple composite of all subject scores was obtained by simply averaging the individual scale scores. This "non-equated" composite would not adjust for group ability and/or subject and test differences. The results of these analyses are presented in the next section.

As outlined in the previous section, the equipercentile equating allows for comparison among three key derived or total score composites. First is the total score for AIEEE (BENG_TOT). This subject has been named in these analyses as AIEEE. Second is the composite score obtained by averaging scores on each subject examination (See Table 1) across **all** of India, after converting these scores to their equivalent AIEEE score via the equipercentile linking procedure described in the previous section. This variable has been named AIEEE-EC for "AIEEE Equated Composite". Finally, there is the simple composite score obtained by averaging the individual subject scores (see Table 1) without any linking or equating adjustment. This variable has been named SS-C for "Scale Score – Composite". Table 2 presents the summary descriptive statistics for these composites.

TABLE 2

Descriptive Statistics

	AIEEE	**AIEEE-EC**	**SS-C**
Mean	58.0295	61.1826	67.9993
SD	50.0268	46.3334	15.6629
Min	-51	-29	9
Max	345	330	99
N-count	185123	184947	185123

Table 2 reveals the similarities as expected between AIEEE and AIEEE-EC. Also seen in this table is the restriction of range imposed on the AIEEE-EC (minimum and maximum scores less than those seen for AIEEE) most likely to do with the 0-100 originating scale associated with each test linked to the much larger AIEEE scale which ranges from -51 to 345.

Table 3 provides the simple inter-correlations amongst the three scores

TABLE 3

Simple Correlations

	AIEEE	AIEEE-EC	SS-C
AIEEE	1.0000		
AIEEE-EC	0.5661	1.0000	
SS-C	0.5625	0.8572	1.0000
	N-count =	184,947	

One result that seems relatively surprising on first review is the strength of the correlations between the two derived composite scores, AIEEE-EC and SS-C relative to the correlations between AIEEE and the composite scores AIEEE-EC and SS-C. Why would AIEEE-EC correlate higher with SS-C than with what it is supposed to be equivalent to, namely AIEEE? This is likely an artifact of the data and has to do with the fact that the rank orderings of the two derived composite scores were predestined to stay the same given the linking procedure used in this paper. The equating adjustment used to obtain AIEEE-EC relied on the rank ordering or percentile rank of each subject scaled score. As such, the composite generated from AIEEE-EC should indeed rank students in a similar way as the simple composite SS-C since both rely on the rank ordering of the students taking each examination.

The relatively weak correlation between AIEEE and AIEEE-EC; and,AIEEE and SS-C testifies that the rank ordering of students based only on AIEEE is different from the rank ordering of students on the two composite scales (i.e. AIEEE-EC and SS-C). This is likely due to the fact that the various subjects will each have different relationships to AIEEE with subjects such as physics and mathematics likely to be more strongly related while areas like psychology and languages less so.

Table 4shows a misclassification table. It is used to summarise the differences that occur from producing a rank order of achievement based on total scaled scores (AIEEE-EC) and a rank order of achievement based on raw scores (SS-C). For the purposes of this illustrative example, it uses the top 10^{th} percentile as the cut-off score. IN on the ***scaled score*** means that the students are in the top 10% of students on the AIEEE-EC and OUT means that they are below the top 10 percent, Similarly, IN on the ***raw score*** means that the students are in the top 10% of students on the SS-C and out means that they are below the top 10%.

TABLE 4

Misclassification Table Showing the Differences in Results Before and after Scaling

		SCALED SCORES		
		IN	**OUT**	**TOTAL**
RAW SCORE	**IN**	10,380 (5.6%)	7,475 (4.0%)	17,853 (9.6%)
	OUT	8471 (4.6%)	158,797 (85.8%)	167,268 (90.4%)
	TOTAL	18,851 (10.2%)	166,272 (89.8%)	185,123 (100.00%)

Table 4 shows those students that are in the top 10% of students on both scaled scores and raw sores (10,380 or 9.6% of the sample); those who are in the top 10 % on neither score (158,797 or 85.8%); those who are in the top 10% on the raw score and not in the top 10% on the scaled score(7,475 or 4.0%); and, those who are in the top 10% on the scaled score but not the raw score (8,471 or 4.6%). It is the last group of students who would be disadvantage unfairly if scaling were **not** carried out.

It is important to stress the following when interpreting these results. Firstly, when producing a single rank order of merit based on the aggregate of subject scores in which not all students have attempted the same subjects then it is essential that scaling (or equating) be carried out to produce scores which are more valid then those produced by just aggregating the raw scores.

Secondly, this example only uses the data from the CBSE in 2012 where in most cases every student has done at least the BENG_TOT; Chemistry; Mathematics; Physics; and, English Foundation. When the results from different Boards are included then the impact of scaling will become much more significant in terms of its impact on students.

Some Potential Problems with Equipercentile Linking (Scaling)

One of the main concerns in using this procedure to render the scores across subjects and Examination Boards comparable is that the anchor test (AIEEE in this case) irrespective of which one is used will be differentially valid for different subjects. That is, it will correlate better with some subjects than others. The higher the correlation between the scores on the anchor test and the cores on the subject the more valid it is to equate scores in the subject.

McGaw (1983) summarises the problem with using a test like the AIEEE or JEE (Main) as the anchor test for equating different subject tests as follows:

> In each rescaling ASAT (anchor test in Australia) is used essentially to identify the characteristic of the candidates enrolled in order to determine how they stand in relation to the other students who might have enrolled. In a subject like chemistry where the correlation with ASAT is high, the ASAT scores of the students enrolled give a reasonable indication of their relative standing in chemistry in a

> population where all students took chemistry. In a case like economics or French, ASAT provides a less valid indication of the selectiveness of the students enrolled. ASAT is thus a more valid rescaling variable for chemistry than for economics or French.
>
> (McGaw, 1983, page 9)

A further consequence of the lack of homogeneity in the inter-subject correlations is that the aggregates constructed from different combinations of subjects will have will have substantial differences. When scores are aggregated to form averages the variances of the sum will be the sum of the variances of the individual subjects that comprise the aggregate plus twice the covariance between each pair of subjects. Aggregates comprised of subjects that have relatively high inter-correlations will have greater variance. The effect will be that the results for persons above the mean will be pushed higher above the mean of the aggregate. Those students who take combinations of subjects which do not have a high inter-correlation will not be pushed as high.

In the competition for entry into it's the problem is less likely to be an issue because the applicants will be inclined to include subjects like subjects in their aggregates.

In educational measurement there is a need to distinguish between actually aggregating scores as has been carried out in this option and the process of measuring. The next option provides an alternative that uses one of the family of Rasch Models to construct a measurement scale which can then be used to measure the property of interest.

One of the advantages of using such a measurement model is that it provides an explicit framework for evaluating the validity of equating any two tests. Once it has been established that equating has been conducted at the level of tests, and then the achievement of the students can be located on the same scale. Together with the achievement measure the model provides an estimate of standard error and an index of fit between the person's profile of results and the model.

Option 4

The fourth option is to develop a measurement scale using an extension of the Simple Logistic Model (SLM) of Rasch. The Extended Logistic Model (ELM) is a generalisation of the SLM for cases where the items have more than 2 ordered response categories. It evolves from an elaboration of Rasch's generalised model as a consequence of substantial work done by Andersen (1973 and 1977) and Andrich (1978). This study uses the ELM where the subjects are treated as polytomous items with scores that range from 0 to 100.

$$\Pr\{X = x; \beta, \delta, \kappa\} = \exp\left[\kappa_x + x(\beta - \delta)\right]/\gamma$$

where β is the student achievement;
δ is the subject difficulty; and

$$\gamma = \sum_{k=0}^{m} \exp\left(\kappa_x + k(\beta - \delta)\right)$$

In this option, the model is used to first create a scale and then measure performance of students against the scale.

The intention in this paper is to demonstrate how the model can be used to generate aggregate scores for students which take account of the relative difficulties of the different subjects and enables their direct comparison on a single rank order of merit that can then be used to directly compare performance for the purposes of competitive entry into IITs.

It is not the intention of this paper to go into detail regarding the model. However, Tognolini (1989) and Tognolini and Andrich (1995, 1996) have demonstrated how the model can be used to generate scores for entry into universities.

In order for the scale to be generated there has to be something that the students have done in common. In 2012 the AIEEE was done by most students wanting to apply to the IITs. In future this will be replaced by the JEE (Main).

Illustrative Example

Data from the CBSE 2012 examination results were used to demonstrate the model. The data consisted of results for 185,123 students in 68 subjects (BENG_TOT is the same as the AIEEE) (See Table 5).

It can be seen from Table 1 that a number of subjects had no enrolments for the sample of students chosen. These subjects were removed for the illustrative example. Other subjects with a very small number of candidates were also removed for the study.

The final sample consisted of 31 subjects (See Table 6).

TABLE 5

CBSE Sample of Subjects and Student Numbers for 2012

SUBJECT	No. of students
BENG_TOT	185123
Mathematics	184947
Physics	185123
Chemistry	185123
English Elective (001)	192
Hindi Elective (002)	1056
Urdu Elective (003)	36
Sanskrit Elective (022)	27
History (27)	13
Political Science (28)	41
Geography (29)	318
Economics (30)	4203
Music Car Vocal (031)	0
Music Car Ins (032)	0
Music Hind Vocal (34)	5176
Music Hind Ins Mel (35)	308
Music Hind Ins Per (36)	76
Psychology (037)	164
Sociology (039)	59
Biology (44)	23735
Biotechnology (045)	700
Eng Graphics (046)	2353
Phys Education (048)	91163
Painting (049)	8720
Graphics (50)	8
Sculpture (051)	66
APP Commercial Art (052)	1270
Fashion Studies (053)	56
Business Studies (054)	4
Accountancy (055)	35
Dance Kathak (056)	56
Dance Bhar (057)	3
Dance Odissi (059)	0
Home Science (064)	650
Informatics Practice (065)	9718
Entrepreneurship (066)	274

Multi Media & Web (067)	894

TABLE 5 (continued)

CBSE Sample of Subjects and Student Numbers for 2012

SUBJECT	No. of students
Agriculture (068)	51
Graphic Design (071)	7
Mass Media (072)	18
Computer Science (83)	46152
Functional English (101)	5571
Punjabi (104)	444
Bengali (105)	380
Tamil (106)	38
Telugu (107)	84
Marathi (109)	121
Manipuri (111)	1
Malayalam (112)	155
Oriya (113)	5
Assamese (114)	0
Kannada (115)	187
Arabic (116)	25
Tibetan (117)	14
French (118)	7
German (120)	5
Russian (121)	0
Nepali (124)	44
Limboo (125)	1
Lepcha (126)	0
Bhutia (195)	1
English Core (301)	179346
Hindi Core (302)	20077
Urdu Core (303)	18
Sanskrit Core (322)	1818
Typography (607)	133
Marketing (613)	1
Geo Spatial Technology (740)	32

TABLE 6

CBSE Subjects and Student Numbers used in the Illustrative Example

Serial Number	SUBJECT	No. of students
1	BENG_TOT	185123
2	Mathematics (41)	184947
3	Physics (42)	185123
4	Chemistry (43)	185123
5	English Elective (001)	192
6	Hindi Elective (002)	1056
7	Geography (29)	318
8	Economics (30)	4203
9	Psychology (037)	164
10	Sociology (039)	59
11	Biology (44)	23735
12	Biotechnology (045)	700
13	Eng Graphics (046)	2353
14	Phys Education (048)	91163
15	Painting (049)	8720
16	APP Commercial Art (052)	1270
17	Fashion Studies (053)	56
18	Home Science (064)	650
19	Informatics Practice (065)	9718
20	Entrepreneurship (066)	274
21	Multi Media & Web (067)	894
22	Agriculture (068)	51
23	Computer Science (83)	46152
24	Functional English (101)	5571
25	Punjabi (104)	444
26	Malayalam (112)	155
27	Kannada (115)	187
28	Nepali (124)	44
29	English Core (301)	179346
30	Hindi Core (302)	20077
31	Sanskrit Core (322)	1818

The BENG_TOT (AIEEE) had a range of scores from -51 to 345. For the purposes of this illustrative study students with scores below 0 and above 100 were excluded. In practice these scores would be transformed to a metric where the top score was aligned to 100 and the lowest score was aligned to 0 for the equating exercise.

In addition, for the purpose of this example, a random sample of 80,000 students was selected for the analysis.

Consequently, the example to illustrate the Rasch (IRT) modeling option used a sample of 80,000 students and 31 subjects.

The Rasch Unidimensional Models for Measurement (RUMM) program was used to analyse the data. Table 7 shows the relative difficulty of the subjects in difficulty order (i.e. from the easiest to the most difficult). Table 7 shows that Painting was the most relatively easy subject; followed by Commercial Art and Informatics Practice. The subject AIEEE was the most difficult.

Once the subjects (items) have been calibrated to produce a scale (Tertiary Entrance Scale) then the students can be measured; that is, placed along the measurement scale.The following equation is used to convert the scores on the different subjects into a single measure of person achievement.

$$r_n = \sum_i x_{ni} = \sum_i x_{ni} . P(x_{ni}; \beta_n; \delta_i; \theta_i; \eta_i; \psi_i)$$

can be used to measure the overall achievement, β_n, for person n.

Students can do any number of subjects; any combination of subjects; and, the resulting β_n will be comparable.

Students in the 80,000 sample used in this example sat different numbers of subjects; some sat 7 subjects, some sat 6 and some sat 5. While the resulting β_ns are comparable irrespective of the number of subjects attempted, only those students (who sat 6 subjects (including BENG_TOT) were retained for the comparative stage of the exercise (number of students in final sample is 69,140.

TABLE 7

CBSE Subjects in Difficulty Order (Logits where the larger the value, the more difficult the subject)

Serial Number	SUBJECT	Difficulty[3]
1	Painting (049)	-2.498
2	APP Commercial Art (052)	-2.363
3	Informatics Practice (065)	-0.917
4	Eng Graphics (046)	-0.565
5	Phys Education (048)	-0.509
6	Agriculture (068)	-0.489
7	Hindi Elective (002)	-0.283
8	Hindi Core (302)	-0.255
9	Home Science (064)	-0.225
10	English Core (301)	-0.025
11	Sociology (039)	-0.009
12	Punjabi (104)	0.093
13	Geography (29)	0.098
14	Nepali (124)	0.165
15	Psychology (037)	0.252
16	Chemistry (43)	0.262
17	Sanskrit Core (322)	0.266
18	Entrepreneurship (066)	0.268
19	Multi Media & Web (067)	0.302
20	Malayalam (112)	0.31
21	Biology (44)	0.336
22	Fashion Studies (053)	0.338
23	Physics (42)	0.352
24	English Elective (001)	0.376
25	Computer Science (83)	0.428
26	Biotechnology (045)	0.506
27	Mathematics (41)	0.539
28	Functional English (101)	0.6
29	Economics (30)	0.645
30	Kannada (115)	0.649
31	BENG_TOT	1.355

[3]These are expressed in logits (i.e. Logarithmic Units). The more positive the value, the harder the subject.

The following misclassification table (Table 8) is used to summarise the disparities that occur from producing a rank order of achievement based on total scores and a rank order of achievement based on raw scores.

TABLE 8

Misclassification Table Showing the Differences in Results Before and after Scaling

		SCALED SCORES		
		IN	**OUT**	**TOTAL**
RAW SCORE	**IN**	13,449 (19.5%)	438 (0.6%)	13,887 (20.0%)
	OUT	480 (0.6%)	54,773 (79.2%)	55,253 (80.0%)
	TOTAL	13,929 (20.1%)	55,211 (79.9%)	69,140 (100.00%)

Table 8 shows those students that are in the top 20% of students on both scaled scores and raw sores (13,449 or 19.5% of the sample); those who are in the top 20 % on neither score (54,773 or 79.2%); those who are in the top 20% on the raw score and not in the top 20% on the scaled score 438 or 0.6%); and, those who are in the top 20% on the scaled score but not the raw score (480 or 0.6%). It is the last group of students who would be disadvantage unfairly if scaling were **not** carried out.

While it appears as though less than 1% of students would be disadvantaged by the introduction of scaling it is important to stress the following when interpreting these results. Firstly, it is essential, when producing a single rank order of merit based on the aggregate of subject scores in which not all students have attempted the same subjects that scaling (or equating) be carried out to produce scores which are more valid then those produced by just aggregating the scores. That is, it is the right thing to do.

Secondly, this example only uses the data from the CBSE in 2012 where in most cases every student has done at least the BENG_TOT; Chemistry; Mathematics; Physics; and, English Foundation. The only variation occurs really in the inclusion of the sixth subject. When the results from different Boards are included then the impact of scaling will become much more significant in terms of its impact on students.

Thirdly, one of the main advantages of using a measurement model to govern the scaling process is that the measurement model provides a means of explicitly evaluating the validity of the scaling process. This has not been done at this stage as the intention of this paper is to show the effects that scaling has on the rank ordering of the top 20% of the students based upon their subject scores.

Some Potential Problems with Rasch Equating

One of the main issues with equating using the Rasch Model is that generally, with such large numbers, the data do not fit the model. This could be a potential criticism that can be addressed; not by getting the data to fit the model but showing that for the purposes of equating the Model is robust to the magnitude of the variation that

generally occurs. At the same time, with feedback from the process to the examiners who set the subject papers, it is anticipated that the fit to the model will improve over time.

A second, more practical issue, is that the current programs used for Rasch modeling of the type advanced here are generally not built to handle the volumes of data that will be expected when all students are included in the scaling process. Once again this is a problem that can be readily addressed as it primarily relates to the amount of memory allocated within the program for the analysis.

6.0 Conclusion

Combining together information of disparate background has confronted researchers, statisticians and lay people alike. There are a number of methods that will enable the aggregation of data across variables to generate a combined score which has meaning in some augmented variable. In the United States for example, the Dow Jones Industrial Average (DJIA) is an indicator of the status of one of the North American stock markets. This averaged is a composite score that averages many different stock holdings together.

Similarly, in education, makers of achievement and aptitude assessments have traditionally created composite scores that combine disparate pieces of information. The Stanford Achievement Test Series, Tenth Edition (Pearson, 2012) for example creates a total test score by combining areas of mathematics, reading, science and history. The ACT Assessment (2009) used in the US is a premier college entrance examination.

The ACT Composite is made up of English, mathematics, Reading, and science measures. In Australia, most states and territories use scaling or equating of one form or another to solve the problem of producing a single rank order for tertiary entrance when students have taken different combinations of subjects.

Hence, while at first reading it seems illogical to combine scores from say mathematics with language, it is often done and done so to provide a higher level variable representing a more generalised ability or skill. While the principles of statistical test form equating and scaling have strong requirements when combining or linking scales via equating, this paper investigated some options that might be considered when combining subjects to form a composite that can be used to rank order students across the country for the purposes of entry into IITs that will ensure that students will not be advantaged or disadvantaged by the subjects they choose or the Board they have chosen to accredit their performance.

While the procedures do have their limitations, the question becomes one of magnitude of error - namely, do decisions that result from linking disparate scores become more accurate after scaling than they would be if the scores were not scaled at all. The answer to this must be yes.

The results of this investigation are relatively straight-forward. Firstly, it seems that the application of the Rasch Scaling model sufficiently reproduce a rank ordering of subjects in terms of difficulty that are intrinsically rationally valid. That is, there seems very little incongruence with a priori expectations of the ordering of difficulty and what was discovered.

Secondly, the equipercentile linking showed the various relationships between composite scores and how, in the aggregate, the power of the subjects taken by most

students (the languages like MRK_041, MRK_042, MRK_043) as well as mathematics, physics and chemistry (MRK16, MRK17, MRK18) are likely dominating the composite score making the composite seem more similar to AIEEE than it should. This suggests that a more extreme situation (where say the five easiest subjects taken by the most liberal Boards are compared against students taking the five most difficult subjects offered by the most conservative Boards) might reveal a more likely potential for bias for composite scores not using linking to adjust for differences in group ability and subject difficulty

Third, the correlations after the equipercentile linking suggest that there are substantial differences in the ordering of subjects using composite scores than in using the AIEEE total score only. This is a serious fairness concern with the current desire to mix subjects and Board results into a single tertiary determination.

Finally, the analyses have shown that, with relative ease and exploiting the fact that most students will always sit for the AIEEE exam, little additional effort or time will be lost in providing some linking adjustment before reporting scores.

7.0 References

ACT. (2009). ACT's College Readiness System: Meeting the challenge of a changing world. Iowa City, IA: ACT.

Albano, A. (2011). *Statistical Methods for Test Score Equating*. R Package Version 1.1-4. Installed from web http://www.r-project.org.

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 of Education.

Bèguin, Anton A., (2000). Robustness of Equating High-Stakes Tests. Unpublished Dissertation, UniversiteitTwente, The Netherlands.

Braun, H. I., & Holland, P. W. (1993). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In: P.W.Holland,& H. Wainer (Eds.), Differential Item Functioning (pp.25-29). Hillsdale, NJ: Erlbaum.

Dorans, N. J. (1990). Equating methods and sampling designs. Applied Measurement in Education, 3, 3-17.

Gulliksen, H. (1950). Theory of mental tests. New York: Wiley.

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

Kolen, M. J.(1981).Comparison of traditional and item response theory methods for equating tests. Journal of Educational Measurement, 18, 1-11.

Kolen, M. J. (1985). Standard errors of Tucker equating.Applied Psychological Measurement, 9, 209-223.

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

Kolen, M. J. & Brennan, R. L. (2004).*Test Equating, Scaling, and Linking*. (2nd ed.), New York: Springer.

Lord, Frederic M. (1977). Some item analysis and test theory for a system of computer-assisted test construction for individualized instruction.Applied Psychological Measurement, 1, 447-455.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Marco, G. L., Petersen, N. S. & Stewart, E. E. (1983). A test of the adequacy of curvilinear score equating models. In D. Weiss (Ed.), New horizons in testing (pp.147-176). New York: Academic.

Mislevy, R. J. (1992).Linking educational assessments: Concepts, issues, methods, and prospects. Princeton, NJ: ETS Policy Information Center.

Petersen, N. S., Kolen, M. J.& Hoover, H. D. (1989).Scaling, norming and equating. In R.L.Linn (Ed.), Educational Measurement (3rd ed.,pp.221-262). New York: American Council on Education and Macmillan.

Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests.(Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Rasch, G. (1968). A Mathematical Theory of objectivity and Its Consequences for Model Construction. European Meeting on Statistics, Econometric and Management Sciences. Amsterdam 2-7 September 1968. pp. 31.

Rasch, G. (1977) On specific objectivity: an attempt at formalizing the request for generality and validity of scientific statements. Danish Yearbook of Philosophy 14: 58-94.)

Thurstone L. L. 1959. The Measurement of Values. Chicago: University of Chicago Press

Whitely, S. E. and Dawis, R. V. (1974), The nature of objectivity with the Rasch model. Journal of Educational Measurement, 11: 163–178.