Most Measures are Still Uncommon: The Need for Comparability

Jon S. Twing, Ph.D.
Managing Director: Pearson Assessment Centre

James S. Tognolini, Ph.D.
Senior Vice President (Research and Assessment)
Pearson Assessment Centre

This paper makes explicit the desire and need to link disparate assessments in large-scale assessments. Furthermore, it shows that well established and well-defended studies (like the Anchor Test Study and those associated with the Voluntary National Tests in the US), disagree regarding the viability, validity and practicality of linking disparate assessments. We point out what is now a common procedure in Australia for linking disparate statewide assessments. Furthermore, we present a scenario in India in which we make the argument that such a linking is better than no linking at all. We demonstrate outcomes of one implementation of this linking and share the results both good and bad. We then suggest more research to answer some of the important questions associated with ultimate policy decisions related to this research.

Key words: Equating, selection, scaling and psychometrics

Introduction

Measurement experts and test builders have for a long time desired an efficient means to link and compare scores for students on different assessments typically given at different points in time for different purposes. In the United States, for example, World War I required the rapid selection and identification of candidates for service and assignment of these recruits to functions based on their skills and aptitudes. To accomplish this, psychologists developed the Army Alpha Test, a written test to sort recruits in such a way. Immediately, however, it was clear that not all recruits could read. Hence, the development teams built an "equivalent" test that was performance-based and known as the Army Beta Test (McGuire, 1994).

The Anchor Test Study (1972)

As the psychologists involved with the Army Alpha and Army Beta tests extended the concept of group administered testing following World War I, it is not hard to speculate that they faced countless issues with score comparability. In fact, Cureton in 1941 (Cureton, 1941) may have been the first to argue for a more comprehensive approach to score comparability or what Linn (Linn, 1975) referred to as Cureton's "original plea for an anchor test study."

What Cureton was looking for was a way to compare individual student scores from one test to another. Presumably, this would allow a psychologist, licensing agency or school to evaluate patients, clients or students with the most efficiency by using all available

assessment data without having to re-administer existing assessments or replicate evaluations.

In the U.S., Cureton got his "anchor test study" via the federally funded Anchor Test Study (ATS) of 1972 (Bianchini & Loret, 1972a; Bianchini & Loret, 1972b). ATS had two major goals: generating comparable individual student scores in reading across tests, and; providing new national norms for each of the seven tests selected (Linn, 1975). For our purposes, we will focus only on the comparability of scores from the ATS. Many people have argued that because of differences in content, format, and specifications linking or equating such tests would not likely to be successful (Angoff, 1964; Lindquist, 1964). In even a restricted definition, two measures to be linked to should measure the same thing—and this is where the arguments begin. Generalized constructs like "mathematics" or "reading" seem to be universal until you look at just what and how is being measured.

Linn (1975, pg. 207) calls this equating "quite satisfactory for most practical purposes." The procedure used to perform the linking for the ATS was an equipercentile method pooled for counter-balanced orders of administrations and, according to Linn was found to be "most satisfactory".

So, it would seem that Cureton (1941) would be satisfied in that his desire for an "anchor test study" was indeed carried out via the ATS. Not only that, but that it was successful based on estimates of the resulting equating error.

The success of the ATS led Linn to believe that such comparability between disparate assessments was possible in at least reading and, that this would add value to the ability of stakeholders to have more flexibility in decision making.

Linking for the Voluntary National Tests (1999)

Given the success of the Anchor Test Study in 1972, it is with some curiosity that a very similar set of questions and resulting inquiry in 1997 in the U.S. led to exactly the opposite conclusions. As part of the Voluntary National Tests, Congress commissioned the National Research Council to investigate the possibility of linking individual student scores from various state assessments to the "Nations Report Card" known as the National Assessment of Educational Progress or NAEP (Feuer, M. J., et. al., 1999). The Committee on Equivalency and Linkage of Educational Tests chose two criteria for the evaluation. First and foremost was the validity of the score interpretations resulting after linking. Second was the "practicality" of the linkage.

Even in 1996, the movement to compare assessment results in the U.S. with other international assessments was in hot debate. The Committee (Feuer, et. al., 1999, pg. 7) argued that one of the motivations for the desire to link disparate assessments in the U.S. was that American educators needed information about not only how students were performing nationally but also how well they were performing when compared internationally.

Given the many ways tests desired to be compared could be different, the Committee (Feuer, et. al., 1999, pp. 91-92) basically concluded that in the US linkages between statewide assessments and/or between statewide assessments and the NAEP assessment simply could not be done.

Presumably, on the heels of the Anchor Test Study (Bianchini & Loret, 1972a; Bianchini & Loret, 1972b), with its results that were seemingly opposite the findings of the Committee on Equivalency and Linkage of Educational Tests (a study that the Committee itself cited as a "model of linkage development" (Feuer, et. al., 1999, pg. 25) the reader should be dismayed and perplexed regarding the viability of linking multiple assessments. The Committee did leave the door open for more investigation claiming that what is not known is the "level of precision needed to make valid inferences about linked tests." (Feuer, et. al., 1999, pg. 93). One may contextualize this to mean that the "invalidity" associated with linking assessments, even ones that have differences in content, reliability, formats, etc., may be "small enough" to justify the linkages depending upon the consequences of other alternatives. Just how important are the differences that would result from links without error to links with these sources of "invalidity" or threats to inferences made from the scores. If these differences are small and unimportant, why not link? These are exactly the questions researchers asked when they linked other such assessments internationally as we will see in the next section of this report.

What the previous section outlined was research done in the U.S. regarding the desire to make scores from disparate assessments comparable. Sometimes the desire for comparability is for efficiency: for example, when a new student comes to a school with a record of past accomplishments (including achievement test scores) that are not the same as those used locally. In such a situation it would be nice to have a "conversion table" to "translate" the scores from the previously used assessment to the current scale of record without having to administer another assessment. Sometimes this desire for comparability is for policy or evaluation. This might occur when a stakeholder in one jurisdiction asks how well their students (or candidates) are performing relative to another jurisdiction. Or, this might be to evaluate specific policy implementations across jurisdictions to see where changes in policy worked and where it did not. Regardless, comparability of assessment results across these jurisdictions is required. Finally, one reason is fairness. Perhaps a student or candidate in one jurisdiction will be competing with a student or candidate from another jurisdiction for enrollment into university. In this situation presumably both candidates will want to be treated fairly in the selection and enrollment process but it is unlikely both candidates will have identical backgrounds—they will likely have taken different courses, received different grades, taken different examinations, etc. How then can a fair comparison be made and a fair selection be facilitated? One way would be to link the assessment scores for these candidates. This seems simple enough but the next section explains just how difficult such a "simple solution" might be.

Linking Assessments for Tertiary Institutions in Australia (1989 – Present)

Tognolini (1989) points out that tertiary institutions in Australia, due primarily to the limited number of seats available for newly entering students, had to rely almost exclusively on an assessment system for selection of students. In this regard, they created a Tertiary Entrance Score (TES) that really, according to Tognolini (1989, pg. 17), "is produced by combining the best assessments a person obtains on a restricted combination of subjects available and acceptable at Year 12". Notice that this implies and, as Tognolini makes explicit later, this means students will not necessarily take the same subset of content. This means that, by definition, almost any attempt at linking in this situation violates the linking criteria presented in the previous section. As such, the question here is not if the linking is viable, but rather, what would result in the fairest comparisons for selection; linking or comparing non-linked scores?

In order to solve this dilemma, Tognolini (1989) used a variation of latent-trait analyses known as the Extended Logistic Model (ELM), which is a generalization of Rasch's simple logistics, or one-parameter "Rasch" model.

A Case Study from India

The 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 (CEE) 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 were implemented in 2013. The latest structure comprises 2 JEE Examinations; JEE (Main) and JEE (Advanced). Candidates wishing to get entry into an Indian Institute of Technology (IIT) have to sit for the (JEE Main). The students who perform best on this examination (top 20%) are then be eligible to sit for the JEE (Advanced) exam.

In addition, the JEE-Main (which was until 2012 known as the All India Engineering Entrance Examination (AIEEE)) is also used for admission to various Central Engineering Institutes other than IITs. The final rank list for entry into these (non-IIT) institutes is prepared by giving 40% weighting to ("normalised") Grade XII Board examination scores and 60% to scores on the JEE-Main.

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 scores 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.

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

The equipercentile method of scaling (variations of which were used for the both Anchor Test Study and the Voluntary National Tests already reviewed) 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. 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. 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.

The equipercentile linking method assumes that the students that have taken the two tests are the same students or at least they are randomly equivalent regarding their test performance. If this is not the case, more advanced equating methods, with additional assumptions must be used.

The basic problem confronted is the following: Changes to the requirements for entry into some institutions from 2013 onwards sees a situation whereby candidates are rank ordered on the basis of a tertiary entrance score which is obtained by giving 60% weighting to ("normalised") Grade XII Board examination scores and 40% to scores on the JEE-Main.

It would be problematic (unfair) to just take the raw scores of students on their examinations done within the various Boards and aggregate them to arrive at a score that can be used to rank-ordering aspiring candidates for competitive entry. This would mean that there has been no attempt to adjust for the relative differences in difficulty of the various subjects taken.

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 generating 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 purely because the subject was inherently easier.

Linking with the Equipercentile method

The equipercentile linking method assumes that in general, scores on different tests cannot be equated using just the mean and standard deviation only (i.e., liner linking). Rather, 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 JEE (Main). Once this scaling has been carried out, the scaled scores for the different subject examinations are comparable.

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, say, 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 from one Board; the Central Board of Secondary Education (CBSE).

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).

Using the linking methodology previously described, each subject was equated to the JEE (Main) total score using the R package "equate" (Albano, 2011).

In this way an equipercentile linking was then performed for all of the subjects such that each subject had a linked JEE (Main) total score equivalent.

A comparison of a composite score obtained from a simple average of these "equated scores" with the empirically obtained JEE (Main) shows the differences possible from decisions resulting from the composite scores relative to decisions made from the use of the JEE (Main) score only.

Table 1
Descriptive Statistics

	JEE (Main)	**JEE-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 1 reveals the similarities as expected between JEE (Main) and JEE-EC. Also seen in this table is the restriction of range imposed on the JEE-EC (minimum and maximum scores less than those seen for JEE (Main)) most likely to do with the 0-100 originating scale associated with each test linked to the much larger JEE (Main) scale which ranges from -51 to 345.

Table 2 provides the simple inter-correlations amongst the three scores.

One result that seems relatively surprising on first review is the strength of the correlations between the two derived composite scores, JEE-EC (Linked Subject Scores) and SS-C (Raw Subject Scores) relative to the correlations between JEE (Main) and the composite scores JEE (Main)-EC and SS-C. Why would JEE-EC correlate higher with SS-C than

with what it is supposed to be equivalent to, namely JEE (Main)? 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 JEE-EC relied on the rank ordering or percentile rank of each subject scaled score. As such, the composite generated from JEE-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.

Table 2

Simple Correlations

	JEE (Main)	**JEE-EC**	**SS-C**
JEE (Main)	1.0000		
JEE-EC	0.5661	1.0000	
SS-C	0.5625	0.8572	1.0000
	N-count =	184,947	

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

Table 3 shows a misclassification table. It is used to summarise the differences that occur from producing a rank order of achievement based on total scaled scores (JEE-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 JEE-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 3

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 3 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 that are more valid then those produced by just aggregating the raw scores.

Secondly, this example only uses the data from 2012 where in most cases every student has done at least the 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.

Limitations of the Study

One of the main concerns in using this procedure to render the scores across subjects and Examination Boards comparable is that the anchor test (JEE (Main) 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 scores on the subject the more valid it is to equate scores in the subject.

McGaw (1983, pg. 9) summarises the problem with using a test like the 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."

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 that do not have a high inter-correlation will not be pushed as high. In the competition for entry into tertiary IT organizations, the problem is less likely to be an issue because the applicants will be inclined to include like subjects in their aggregates.

This study used the only available "common test" (the JEE (Main)) as the anchor to link the board examinations together. Further research is needed to understand if the JEE (Main) shows the properties of a "good" anchor test, as there is some reason to believe it might not (i.e., it is an entrance examination and may not reflect the full range or student response possibilities). It is beyond the scope of this paper to consider this aspect of the anchor test but should be additional research planned.

Finally, while this paper has presented the background and case for performing a statistical link in the context of India, the authors have not investigated the criteria for judging the utility or accuracy of the linking. For example, we have not investigated the properties of the anchor test, we have not evaluated the inter-correlations or the dimensional structure, we have not quantified the error of the linking nor have we compared it to other linking options. Any comprehensive study ultimately impacting policy should include these aspects of evaluation.

Summary

This paper has pointed out that the desire and need to link disparate assessments have been considered for more than a century of our experience in large-scale assessment. Furthermore, we have shown that well established and well-defended studies (like the Anchor Test Study and those associated with the Voluntary National Tests in the US), disagree regarding the viability, validity and practicality of linking disparate assessments. In addition, we presented a scenario in India in which we make the argument that such a linking is better than no linking at all. We demonstrate outcomes of one implementation of this linking and share the results both good and bad. We then suggest more research to answer some of the important questions associated with ultimate policy decisions related to this research.

References

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., (1964). Equating non-parallel tests. *Journal of Educational Measurement*, 1, 11-14.

Beaton, A. E., & Zwick, R., (1990). The effect of changes on the National Assessment: Disentangling the NAEP 1985-1986 reading anomaly. Report No. 17-TR-21, Educational Testing Service, National Assessment of Educational Progress, Princeton, NJ.

Bianchini, J. C., & Loret, P. G., (1972a). Anchor test study. Final report. Project report and volumes 1 through 30, available as ERIC Documents ED 092 601 thorough ED 092 631.

Bianchini, J. C., & Loret, P. G., (1972b). Anchor test study supplement. Final report. Project report and volumes 31 through 33, available as ERIC Documents ED 092 632 thorough ED 092 634.

Cureton, E. E., (1941). Minimum requires in establishing and reporting norms on educational tests. Harvard Educational Review, 11, 287-300.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W. & Hemphill, F. C., (1999). Editors. *Uncommon measures: Equivalence and linkage among educational tests*. Washington, D.C.: National Academy Press.

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

Lindquist, E. F., (1964). Equating non-parallel tests. *Journal of Educational Measurement*, 1, 5-10.

Linn, R. L., (1975). Review. Anchor test study: The long and the short of it. *Journal of Educational Measurement*, (*12*, 3), pp. 201-214.

Lord, F. M. & Novick, M.R., (1968). Statistical theories of mental test scores. Reading, MA: Addison – Wesley.

McGaw, B. (1983). Combining school-based and external assessments of performance at the end of school. Paper presented to the Ninth International Conference of the International Association for Educational Assessment, Blantyre, Malawi, June.

McGuire, F. (1994). Army alpha and beta tests of intelligence. In R. J. Sternberg (Ed.), *Encyclopedia of intelligence*. New York: Macmillan.

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.

Tognolini, J. S., (1989). Psychometric Profiling and Aggregation of Public Examinations at the Level of Test Scores. Dissertation submitted in partial fulfillment of requirements for a degree of Doctor of Philosophy, Murdock University, Western Australia.