Test equating for measuring system progress in longitudinal surveys of student academic achievement

C Nguyen  
P Griffin  
M Wu  
Melbourne Graduate School of Education  
University of Melbourne, Vic 3010 Australia  
Email: cuc@unimelb.edu.au

Abstract. In 2001 and 2007 Vietnam conducted its first and second surveys of student achievement at primary school level in two key subject areas: literacy and mathematics. In order to track the progress of grade 5 students at the system level, the 2001 and 2007 tests were linked using common items. This paper presents the process used to link the two literacy tests. Several issues arose in the equating process of the two tests. This paper discusses these issues and the implications for assessment design and analysis to enable reliable and valid measures of student progress in longitudinal large-scale studies.

1. General description of research questions, objectives and theoretical framework

1.1. Introduction

Monitoring student achievement over time provides a means of tracking system progress over time and assessing possible associations between educational policies and student achievement over the designated period of time. Comparison of student achievement between two points in time is important for evaluating the likely impact of pedagogical interventions on student achievement. There are, however, significant challenges in ensuring that the comparisons are valid, given changes to curriculum and testing procedures. This paper reports on the techniques used to equate the tests of literacy achievement from the Vietnam national surveys of student achievement conducted in 2001 and 2007 in order to place the 2007 cohort on the same scale as the 2001 cohort. The 2001 test has been calibrated, item parameters have been published and the scale has been interpreted [1]. The purpose of this study is to link the 2007 test items to the continuum established in 2001 in order to place the 2007 sample on the same scale as the 2001 sample.

1.2. Framework and research questions

Internationally, there are a number of studies that monitor change in student performance over time, such as the Trends in International Mathematics and Science Study, the Program for International Student Achievement, the United States National Assessment of Educational Progress and the Australian National Testing Program. Each of these studies is characterised by specific approaches to testing, linking and equating. To establish the equivalence of tests over a period of years, many studies link tests across time using a subset of common items. In order to make definitive statements about changes in outcomes, however, the relative difficulty of the items used in the tests should be the same.
To monitor test difficulty, the tests need to be equated, or placed on the same measurement scale, so that they can be directly compared.

The use of item response models, and specifically the Rasch (1960) model, provides one way of achieving this. Item response modelling provides estimates of the difficulty of the items and the abilities of the test-takers on the same measurement scale. This process is called calibration, and is a prerequisite for equating. Three assumptions are required by the Rasch model [2]. First, the construct or variable being measured has a single dimension – that is, only one trait or dimension governs the responses of persons to the items. This refers to the fit of the items when adding a new set of items to the existing set of items. Second, person and item measures can be placed on the same scale. The difference between item difficulties should be stable and free from sample of people who are tested [3]. Therefore the common items should not demonstrate differential item functioning (DIF) in relation to test times or versions. And in order to meet these requirements, one further assumption is needed. Each item must be independent of all others, so that each can be used as a pointer or indicator of position on the variable being measured. If this is not done and tests are designed in a way that creates local item dependency, as in assessment instruments that include passages of text and also items embedded in those passages, increased linkage errors can result [4, 5]. Monseur & Berezner (2007) pointed out the need for calculating and including equating errors when reporting the trends of educational achievement over time. Failing to recognise linking errors leads to an underestimation of standard errors and thus increases the Type I error rate, thereby resulting in reporting of significant changes in achievement when in fact the changes are not significant [4]. Large-scale surveys of student achievement often describe a continuum of student abilities and report student achievement in terms of level of proficiency. Therefore the descriptions of the derived scales after equating should be compared to the original scale descriptions as a way to examine if the underlying proficiency continuum was the same after equating. It should also be noted that equating usually requires large samples and depends on the equating method and design [6]. The location of test items in a test may impact item difficulties. The same items when moved from the beginning of a test to the end of a test may change their difficulties [7, 8]. It has also been argued that the link items should be spread out to cover different level of a scale continuum. In addition, the number of items must meet a minimum requirement [9].

The research questions to be addressed in this paper are:

- To what extent do the common items of the two tests, 2001 and 2007 Grade 5 reading tests (to be used for calibration), meet the requirement for ‘good link items’ under the Rasch model?
- What is the best method for equating the 2001 and 2007 literacy tests?

2. Methodology

2.1. The data and the sample

The data were sourced from two national studies of student achievement in Vietnam. The first study, in 2001, assessed both students and teachers in mathematics and reading. The second study, in 2007, assessed students only. The multistage sampling process that governed the collection of the achievement sample data is described in detail in the first Vietnam study report on the 2001 survey [1]. The 2001 sample include 72,666 Grade 5 students while the 2007 sample included 59,601 Grade 5 students. The data therefore, meet the requirement of large samples for equating design [6].

The two literacy tests were developed using the same test blueprint [1] which meets the assumption and requirement of the Rasch model for test equating purpose [9]. The 2001 literacy test included 60 items while the 2007 literacy test included 40 items. There were 12 common items between the two tests which was considered small compared to the minimum number of items required suggested by Wu et al. (2016) [8]. The 12 items were embedded in the three passages. This is not an ideal situation
as it may create an issue of local item dependency which in turn can lead to increased linkage errors [4,5]. The locations of items were not the same across the two tests which may impact on the item difficulties and therefore may lead to item differential functioning (DIF). Given the data characteristics, great attention has been given to identifying a suitable method that enhances reliability and validity of the test equating process for this common item design.

2.2. The analysis
Before conducting test equating, the first step was to examine the extent to which the common items met the requirements of good link items. In accordance with the literature review presented in the previous section, the following criteria were checked:

- The extent to which the items fit the Rasch model: It is required that when the two tests were calibrated separately the link items should have good fits, indicating that the link items are in agreement with other items in measuring a single construct of students’ literacy competency. It is also required that the fit measures of the common items, when test equating is conducted, are good.
- The extent to which the link items cover a range of student abilities
- The extent to which the links items behave consistently across time: This means that there is a relative stability of difficulties of the link items across different calibrations.
- Equating error: As the equating error should be included when describing the progress trend from 2001 to 2007, it is expected that the equating error should be relatively small.
- The extent to which the proficiency scale underlying the 2007 test be interpreted in the same way as the 2001 test.
- The extent to which the descriptions of the proficiency levels established in 2001 match the content of the items clustered in the those levels when applying the cut scores used in 2001 to the 2007 continuum.

As a result of the first step, items which did not meet these requirements would normally be removed and a set comprising only link items that had stable difficulties and met the above criteria would be retained. The situation in this study is problematic, as the link items do not meet the requirements of good link items. However, as the number of link items in this study was 12 and did not meet the required number, [8] removing items is not a good solution.

3. Results

3.1. Evaluation of the link items
In order to check if the common items fit well with other items to measure a single domain, the Weighted Mean Squared Residuals (MNSQ) and Weighted t-statistics of the common items in the 2001 and 2007 test calibrations were examined. The MNSQ of the common items in two calibrations were relatively close to 1 which shows that the common items had a good fit. The Weighted t-statistics, however, were statistically significant. When sample sizes are large, t-statistics tend to be significant [8]. In this study the 2001 and 2007 sample sizes were 60000 and 72000 respectively. Therefore the significance of the t-statistics in this case can be ignored. As a result, it can be considered that the common items fit well with other items to measure a single domain.

The difficulties the common items obtained from the separate calibrations of the two tests are presented in Table 1. Column 1 presents the reading passages which are associated with items. Column 2 presents the position of the link items in the 2011 test while column 3 presents the position of the link items in the 2007 test. Column 4 presents the item difficulties of the link items when calibrating 2001 test items alone. Column 5 presents the item difficulties of the link items when calibrating 2007 test items alone. Column 6 presents the adjusted item difficulties of the link items in 2007 onto the 2001 item scale using “a shift” G_0107 which was calculated using the following formula: 

\[ G_{0107} = \text{shift} \]
\[
\sum \left( d_{i01} - d_{i07} \right) / k \text{ where } k \text{ is the number of link items between 2001 and 2007, and } d_{i01} \text{ and } d_{i07} \text{ are the estimated difficulties of the link item } i \text{ in each test [10]. It can be seen from column 6 that the 2007 item difficulties has adjusted to have the same mean with the 2001 item difficulties but still keep their own standard deviation. Column 7 presents the difference in difficulties of the link items between 2001 and 2007 by subtracting the adjusted 2007 item difficulties from the 2001 item difficulties.}
\]

Table 1. The difficulties the common items: free calibration and transformed using shift mean.

| Reading passage | Item position in 2001 test | Item position in 2007 test | 2001 logit from free calibration | 2007 logit from free calibration | 2007 adjusted logit using shift mean | Difference of 2001 and 2007 item difficulties (d_{i01} - d_{i07adj}) |
|-----------------|-----------------------------|-----------------------------|---------------------------------|---------------------------------|--------------------------------------|----------------------------------------------------------|
| (1)             | (2)                         | (3)                         | (4)                             | (5)                             | (6)                                  | (7)                                      |
| A               | 48                          | 9                           | 1.47                            | 0.57                            | 0.88                                 | 0.59                                     |
|                 | 49                          | 10                          | -0.1                            | -0.88                           | -0.57                                | 0.47                                     |
|                 | 50                          | 11                          | 0.21                            | -0.63                           | -0.32                                | 0.53                                     |
|                 | 51                          | 12                          | 0.41                            | -0.27                           | 0.04                                 | 0.37                                     |
| B               | 6                           | 18                          | -1.54                           | -1.38                           | -1.06                                | -0.48                                    |
|                 | 7                           | 19                          | -0.86                           | -0.75                           | -0.43                                | -0.43                                    |
|                 | 8                           | 20                          | -0.26                           | -0.65                           | -0.34                                | 0.08                                     |
|                 | 9                           | 21                          | 0.37                            | 0.20                            | 0.51                                 | -0.14                                    |
|                 | 10                          | 22                          | -0.17                           | 0.07                            | 0.38                                 | -0.55                                    |
| C               | 28                          | 36                          | -1.18                           | -0.84                           | -0.53                                | -0.65                                    |
|                 | 29                          | 37                          | -0.61                           | -0.88                           | -0.56                                | -0.05                                    |
|                 | 30                          | 38                          | 0.8                             | 0.23                            | 0.54                                 | 0.26                                     |
| Mean            |                            |                             | -0.12                           | -0.43                           | -0.12                                | 0.00                                     |
| SD              |                            |                             | 0.85                            | 0.58                            | 0.58                                 | 0.45                                     |

It can be seen from Table 1 that:

- The positions of the common test items in the tests changed between 2001 and 2007. Item positions in the test were found to be associated with a difference in item difficulty. The items become more difficult if they are moved from the beginning of a test to the end of a test. On the other hand items become easier if moved from the end of a test to the beginning of a test. This finding is consistent with what Kingston & Dorans (1984) and Wu et al. (2016) pointed out.
- The set of common items covers a relatively wide range of student abilities.
- The standard division of the item difficulties in 2001 and 2007 are different, providing evidence that item difficulties were not stable across time.
- There are 8 items whose differences in difficulties ranged from 0.37 to 0.59 logit scores. This means that the difficulties of these items are different across two tests and are not qualified as link items for test equating purposes.
There are also other methods of transforming 2007 scale into the same scale of 2001 test calibration: Transform to have the same Mean and SD, using a shift; and Transform to have the same Mean and SD, using regression. Each of these methods identified different subsets of items which were considered as unstable across the two tests.

3.2. Calibration using 12 common items

Given the situation that there was a small number of common items, test equating was conducted with all 12 common items. Four approaches to estimating the link items difficulties were used to equate the two tests:

- A1-Anchoring: the 2001 item difficulties were used to calibrate the 2007 items and estimate 2007 student ability.
- A2-Transform 2007 item difficulties to the same mean and standard deviation of the 2001 item difficulties using a shift mean: In the first stage, the 2007 and 2001 tests were calibrated separately. In the second stage, the item difficulties of 2007 common test items were transformed to have the same mean and standard deviation as the 2001 common items. Other items in the 2007 test were also transformed using the equation for transforming the common items.
- A3-Shift item mean score and no change to standard deviation: In the first stage, the 2007 and 2001 tests were calibrated separately. In the second stage, the item difficulties of 2007 common test items were transformed to have the same mean as the 2001 common items. Other items in the 2007 test were also shifted using the same shift applied for the common items.
- A4-Transform item difficulties to the same mean and standard deviation using regression: In the first stage, the 2007 and 2001 tests were calibrated separately. In the second stage, the item difficulties of the 2007 common test items were transformed to have the same mean and standard deviation as the 2001 test using regression. Other items in the 2007 test were also transformed using the same equation.

The population means which were calculated using the four test equating methods are presented in Table 2.

Table 2. Student ability estimates using different methods – 12 link items.

| Method Description                                      | Minimum | Maximum | Mean | Std. Deviation |
|---------------------------------------------------------|---------|---------|------|----------------|
| A1_Anchor (Anchoring)                                   | 29.93   | 861.99  | 523.66 | 98.66          |
| A2_Transform to have the same Mean and SD, using a shift (shift and scale method) | 29.02   | 899.89  | 542.46 | 108.11         |
| A3_Shift to the same mean(shift method)                | 33.84   | 855.06  | 518.77 | 97.24          |
| A4_Transform to have the same Mean and SD, using regression (scale method) | 32.10   | 879.23  | 532.05 | 102.88         |

It should be noted that the student logit scores have been transformed into the 2001 scale with Mean of 500 and standard deviation of 100.

The linking error was calculated using the following formula

\[ Se_{\text{link}} = \sqrt{\frac{\text{Parameter difference}}{n}} \]  

[4].
where \( n \) is the number of common items.

The equating error was also transformed onto the scale of 2001 reading with a mean of 500 and a standard deviation of 100. As a result, the equating error was 10.3 points, which is 0.1 Standard Deviations (SD) of the student estimates. Here is a summary of the findings from test equating:

- The fit statistics of the common items obtained from the equating exercises were not good, showing that as a result of test equating, the common items may be measuring something different from the other items.
- Depending on the methods of test equating, the gain in literacy scores varied and ranged from 0.19 SD to 0.42 SD.
- Item ordering and location along a literacy continuum was changed dramatically depending on the equating methods, which put the test validity and continuum interpretation under threat.

The test equating using 12 items did not lead to a reliable and valid measures of item difficulties and student abilities. There is a need to try to eliminate the bad link items.

3.3. Identifying poor links and calibration with a smaller number of better link items
The number of items was too small to consider removing poor link items. However, given the situation, an attempt to identify poor link items to remove was made. It was found that, depending on methods of transforming the 2007 common items onto the same scale of 2001 common items, the items flagged as the most unstable items changed. There should be a range of combination of items demonstrating unstable difficulties across the two tests. Technically, all these combinations should be tried as optional and the results should be compared before an optimal solution is found. For this project, it was then decided to combine the expert judgment and psychometrics analysis to come up with a set of four items for test equating with acknowledgement of limitations. A range of test equating was conducted using only four link items; the population mean scores were relatively stable. The test equating error when using the four link items was smaller compared that when using 12 items. The interpretation of the continuum as a result of test equating was compatible to the continuum description established in 2001.

4. Discussion and recommendations
The analysis demonstrates that the design of the 2007 test was not well prepared for equating with the previous test, though an attempt to use the same test blue print and to have common items was made. A strong association found between the change in item positions in the test and the change in item difficulties reinforces the notion that the common items should be placed in similar sections in different tests to avoid item difficulty change between tests [8]. There are multiple methods of identifying items which demonstrate instability in difficulties between the two tests. Each method indicated a different subset of items for removing. Each subset of link items used for test equating and each method lead to different population estimates. While a careful evaluation using the criteria presented above was conducted, the solution used to equate the two tests has a number of disadvantages due to an inappropriate design.

References
[1] World Bank. 2004 Vietnam: reading and mathematics assessment study. Volume 2. Hanoi: World Bank
[2] Rasch G 1960 Probabilistic models for some intelligence and attainment tests. (Copenhagen: Danmarks Paedagogiske Institut).
[3] Wright, B. D. & Masters G. N. 1982 Rating scale analysis. (Chicago: Mesa Press).
[4] Monseur, C., & Berezner, A. (2007). The computation of equating errors in international surveys in education. Journal of Applied Measurement, 8(3), 323–325.
[5] Monseur, C., Sibberns, H., & Hastedt, D. (2008). Linking errors in trend estimation for international surveys in education. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, 1, 113–122.

[6] Kolen, M. J., & Brennan, R. L. (2014). *Test equating, Scaling and Linking. Methods and Practices*. Third edition. (New York: Springer-Verlag).

[7] Kingston, Neal M.; Dorans, Neil J. (1984). *Item location effects and their implications for IRT equating and adaptive testing*. Retrieved from the University of Minnesota Digital Conservancy, http://hdl.handle.net/11299/101880.

[8] Wu, M. Tam H. P. & Jen T. 2016 *Educational Measurement for applied researchers* (Singapore: Springer)

[9] Dorans, N. J., Moses, T., & Eignor, D. (2010). *Principles and practices of test score equating (ETS Research Report No. RR-10-29)*. (Princeton, NJ: ETS).

[10] Wright, B. D., & Stone, M. H. (1979). *Best test design*. (Chicago: MESA Press)