Changes in students' mathematics achievement in Australian lower secondary schools over time: A Rasch Analysis.

Tilahun Mengesha Afrassa

John P Keeves

School of Education

The Flinders University of South Australia

Abstract

This paper aims to analyse and scale mathematics data over time by applying the Rasch model using the QUEST (Adams & Khoo, 1993) computer program. The mathematics achievement of the students is brought to a common scale. This common scale is independent of both the samples of students tested and the samples of items employed. The scale is used to examine the changes in mathematics achievement of students in Australia over time. Conclusions are drawn as to the robustness of the common scale and the changes of students' mathematics achievement over time in Australia.

1. CHANGES IN MATHEMATICS ACHIEVEMENTS OVER TIME

Over the past three decades researchers have shown considerable interest in the study of student achievement in mathematics at all levels across educational systems and over time. Many important considerations can be drawn from various research studies about students' achievement in mathematics over time. Willett (1997, 327) argues that by measuring change over time, it is possible to map phenomena at the heart of the educational enterprise. In addition, he argues that education seeks to enhance learning, and to develop change in achievement, attitudes and values. It is Willett's belief that "only by measuring individual change is it possible to document each person's progress and, consequently, to evaluate the effectiveness of educational systems" (Willett, 1997, 327). Therefore, measuring changes in achievement over time is one of the most important tools for finding ways and means to improve the educational system of a country.

Since, Australia participated in the 1964, 1978 and 1994 International Mathematics Studies, it should be possible to examine the mathematics achievement differences over time across the 30-year time period.

Therefore, the purpose of this study is to investigate changes in achievement in mathematics of Australian lower secondary school students between 1964, 1978 and 1994.

In this paper the results of the Rasch analyses of the mathematics

achievement of the 1964, 1978 and 1994 Australian students who participated in FIMS, SIMS and TIMS are presented and discussed. The paper is divided into seven sections. The sampling procedures used on the three occasions are presented in the first section, while the second section examines the measurement procedures employed in the study. The third section considers the statistical procedures applied in the calibration and scoring of the mathematics tests. The fourth section assesses whether or not the mathematics items administered in the studies fit the Rasch model. Section five discusses the equating procedures used in the study. The comparisons of the achievement of FIMS, SIMS and TIMS students are presented in the next section. The last section of the paper examines the findings and conclusions drawn from the study.

1.1. Sampling procedure

Table 1 shows the Target Populations of the three international studies conducted in Australia. In the First International Mathematics Study (FIMS), conducted in 1964, two groups of students participated, 13-year-old students in Years 7, 8 and 9 and students in Year 8 of schooling. The total number of students taking part was 4320 (see Table 1).

In the first study only government schools in New South Wales (NSW), Victoria (VIC), Queensland (QLD), Western Australia (WA) and Tasmania (TAS) participated. In the Second International Mathematics Study

(SIMS), which was administered in 1978, nongovernment schools and the Australian Capital Territory (ACT) and South Australia (SA) were also involved as well as those states that participated in FIMS. Thus in SIMS government and nongovernment school students in six states and one territory were involved. The total number of participants was 5120 students (see Table 1).

Meanwhile, in the Third International Mathematics Study (TIMS), which was conducted in 1994, government and nongovernment school students in all states and territories including Northern Territory were involved. The total number of students tested was 12852 (see Table 1).

In 1964 and 1978 the samples were age samples and included students from Years 7, 8 and 9 in all participating states and territory, while in TIMS the samples were grade samples drawn from Years 7 and 8 or Years 8 and 9. In ACT, NSW, VIC and TAS Years 7 and 8 students were selected while in QLD, SA, WA and NT samples were drawn from Years 8 and 9.

Therefore, to make the most meaningful possible comparison of mathematics achievement over time by using the 1964, 1978 and 1994 data sets, the following steps were taken.

Table 1:- Target populations in FIMS, SIMS and TIMS

The 1964 students were divided into two groups 13-year-old students in

one group and all Year 8 students including 13-year-old students at

that year level as the second group since in addition to an age sample a grade sample had also been drawn. It is important to observe that 13-year-old students in Year 8 were considered as members of both groups. In the first group students were chosen for their age and in the second group for their year level. The 1978 students were chosen as an age sample and included students from both government or non-government schools. In order to make meaningful comparisons between the 1978 sample comparable and the 1964 sample, students from non-government schools in all participating states and all students from SA and ACT were excluded from the analyses presented in this paper, although Rosier (1980) and Moss (1982) considered both the total student groups and the restricted government school student groups drawn from only five states.

Meanwhile, in TIMS the only common sample for all states and territories was Year 8 students. In order to make the TIMS samples comparable with the FIMS samples, only Year 8 government school students in the five states that participated in FIMS are considered as the TIMS data set in this study. All non-government school students in the five states and all students in SA, ACT and NT are excluded from the analyses presented in this study, although they have been considered in the recent report by Lokan, et al.(1996).

After excluding schools and the states and territories that did not participate in the 1964 study, two sub-populations of students were identified for comparison between occasions. The two groups were 13-year-old students in FIMSA and SIMS: all were 13-year-old students and were distributed across Years 7, 8 and 9 on both occasions. Hence, these two groups of 13-year-old students were considered to be comparable for the examination of achievement over time, between 1964 and 1978. Whereas for the comparison between FIMS and TIMS the other sub-populations consisted of 1964 and 1994 Year 8 students. Students in both groups were at the same year level, although there were differences in the ages between these groups which were tested on the two occasions. Hence, the comparisons in this study are between 13-year-old students in FIMSA and SIMS on the one hand, and FIMSB and TIMS Year 8 students on the other.

1. 2. Measurement procedures employed in the study

In this study the procedures employed to measure mathematics achievement level of students on the three occasions involved the use of the Rasch model to scale students' responses to the mathematics test items. The tests included both multiple choice and constructed response items, and in the 1994 testing program both dichotomous and polychotomous scoring procedures were employed for the constructed response items.

1. 2. 1. Use of Rasch model

The Rasch model has been shown to be the most robust of the item

response models (Sontag, 1984), and was used in this study primarily to equate students' performance in mathematics on a common scale for the Australian investigations conducted in FIMS, undertaken in 1964, SIMS, carried out in 1978 and TIMS, which was conducted in 1994.

1. 2. 1. 1 Unidimensionality

In order to employ the Rasch model for calibrating the items in the mathematics tests it was necessary to examine whether or not the items were unidimensional since the unidimensionality of items is one of the requirements for the use of the Rasch model (Hambleton and Cook, 1977; Anderson, 1994). If the items were found not to satisfy the condition of unidimensionality, it would not be possible to employ the Rasch procedures in the calibration of the tests. Hence, a literature search was undertaken to examine whether or not the test developers had examined the dimensionality of these items, and to the investigators' knowledge the items had at no time been examined for unidimensionality, although Peaker (1969) had considered how the part scores should be weighted.

Consequently, confirmatory factor analysis procedures were employed to test the unidimensionality of the mathematics test items. Confirmatory factor analysis is a statistical procedure employed for investigating relations between a set of observed variables and the underlying latent variables (Byrne, 1989; Kim & Mueller, 1978a, 1978b; Long, 1983; Spearritt, 1994). Thus, confirmatory factor analysis assumes that the observed variables are derived from some underlying source variables

(Kim & Mueller, 1978a). Factor analysis may also be used as an appropriate method for determining the minimum number of hypothetical variables that would account for the observed covariation, and thus as a means of exploring the data for possible data reduction (Kim & Mueller, 1978a). However, one of the main purposes of confirmatory factor analysis is to examine and test the common underlying dimensions associated with a number of observed variables.

The results of the confirmatory factor analyses of FIMS and SIMS data sets revealed that a nested model in which the mathematics items were assigned to three specific correlated first-order factors of Arithmetic, Algebra and Geometry as well as a general higher order factor, which was labelled as Mathematics provided the best fitting model. In addition, in the confirmatory factor analyses undertaken, no evidence was found to reject the assumption of the existence of one general factor involved in the mathematics tests, in so far as in the nested model the Mathematics factor extracted more of the total variance than did the specific first-order factors taken together. Therefore, the mathematics test items in the FIMS and SIMS studies are considered to satisfy the requirement of unidimensionality. The item cluster-based design procedure (Adams and Gonzalez, 1996) employed in the construction of the TIMS data sets would seem to preclude the use of confirmatory factor analysis to test the unidimensionality of the TIMS data set.

1. 3. The statistical procedures employed in the study

In this section the statistical procedures employed in the study are discussed.

1. 3. 1. Effect size

In this paper both the standardized effect size and the magnitude of effect on the calibrated scales are used to examine the level of practical significance of the differences between FIMS, SIMS and TIMS in mathematics achievements over time.

In this study effect size values less than 0.20 are considered as trivial, while values between 0.20 and 0.50 are considered as small. Furthermore, effect size values between 0.50 and 0.80 are taken as moderate and values above 0.80 are treated as large (Cohen, 1991; Keeves, 1992).

1. 3. 2. Growth between grade levels

It is possible since the TIMS project tested in two adjacent grades to estimate the gain between the lower grade and the upper grade for the Australian sample and thus to interpret the calibrated effect size in terms of a year of mathematics learning at the lower secondary school level. The present study revealed that the growth in achievement per year in mathematics achievement in Australian lower secondary schools was 37 centilogits. This value is equivalent to an effect size of 0.30. Keeves (1992) has indicated that an effect size of 0.30 was also found to be equivalent to a year's learning in science at the lower secondary school level . Therefore, this additional information allows the

differences between the achievement level of the different groups to be interpreted in terms of practical significance rather than depending solely on statistical significance.

1. 3. 3. The t -test

In order to determine the level of statistical significance between the mean scores on FIMS, SIMS and TIMS in mathematics achievements a t statistic was calculated, which took into account errors from three scores: (a) sampling error, (b) errors of calibration, and (c) equating error. Further comment on the estimation of these errors is given in Appendix A.

1. 3. 4. Treatment of omits and non-responses

Issues regarding the occurrence and handling of missing data in achievement tests of the kind employed in these three mathematics studies were considered. However, the results of the Rasch analyses did not show marked differences between ignoring the missing data or treating the missing data as wrong during the calibration and scoring. Therefore, for both calibration and scoring purposes it was decided to treat the missing data as wrong. While all the items in the mathematics test were employed for scoring purposes, for calibration purposes only those items that fitted the Rasch scale were considered. The main justification for the use of these procedures would seem to lie in the greater number of misfitting items when the procedure that involved the ignoring of the missing data was tested with the SIMS data. Consequently, the procedure that involved treating missing data as

wrong was chosen for this study.

1. 3. 5. Treatment of zero and perfect scores

The QUEST computer program (Adams and Khoo, 1993) by default does not process cases with perfect and zero scores, because both groups do not provide information for the calibration of the scale. Cases with perfect scores are those cases who provided correct responses for all the items, while cases with zero scores are those cases who provided wrong responses for all items. Hence, in order to include those cases with perfect and zero scores in the calculation of the mean and standard deviation of the mathematics achievement test scores for each student sample, the values of perfect and zero scores were calculated by extrapolation from the logit table produced by the QUEST computer program (see Appendix B). Subsequently, The SPSS 6.1. (Norusis & SPSS Inc, 1990) computer program was used to calculate the case estimate mean scores and standard deviations with appropriate weights, and the WesVarPC 2.11 (Brick, Broene, James and Severynse, 1997) computer program was employed for calculating the standard error of the mean values, again with appropriate weighting of the data, and with allowance made for the fact that all samples were of a stratified cluster sample design.

1. 3. 6. Developing a common mathematics scale

The calibration of the mathematics data permitted a scale to be constructed that extended across the three groups, namely FIMS, SIMS and TIMS students on the mathematics scale. The fixed point of the

scale was set at 500 with one logit, the natural metric of the scale, being set at 100 units. The fixed point of the scale, namely 500 was taken as the mean of the difficulty level of the calibrated items in the FIMS test administered in 1964. The mathematics scale constructed in this way for all different sample groups of students in FIMS, SIMS and TIMS is presented in Figures 1, 2 and 3, with 100 scale units (centilogits) being equivalent to one logit.

1. 3. 7. Conclusion

In the last two sections the scaling and the statistical procedures employed in the study are discussed. The Rasch model was the major scaling procedure employed. The effect size and the t-test were employed for comparing the mean values of different groups of students.

With respect to the missing data a decision was made, from a study of the results of the Rasch analyses, for both calibration and scoring purposes to treat the missing data as wrong, while for calibration purposes only those items that fitted the Rasch scale were employed.

1. 4. Rasch Analysis

Three groups of students namely FIMS (4320), SIMS (5120) and TIMS (7926) were employed in the calibration and scoring analyses. The necessary requirement for calibration in Rasch scaling is that the items and persons must fit the Rasch scale. Items and persons that do not fit the scale must be deleted in calibration. In order to examine whether or not the items and persons fitted the scale, it was also

important to evaluate both the item fit statistics and the person fit statistics. The results of these analyses are presented below.

1. 4. 1. Item fit statistics

One of the key item fit statistics is the infit mean square (INFIT MNSQ). The infit mean square measures the consistency of fit of the students to the item characteristic curve for each item with weighted consideration given to those persons close to the 0.5 probability level. The acceptable range of the infit mean square statistic for each item in this study was taken to be from 0.77 to 1.30 (Adams and Khoo, 1993). Values outside this acceptable range, that is above 1.30 indicate that these items do not discriminate well, and below 0.77 the items provide redundant information. Hence, consideration must be given to excluding those items that are outside this range. In calibration, items that do not fit the Rasch model and which are outside of the acceptable range must be deleted from the analysis (Rentz and Bashaw, 1975; Wright and Stone, 1979: Kolen and Whitney, 1981; Smith and Kramer, 1992). Hence, in FIMS two items (Items 13 and 29), in SIMS two items (Items 21 and 29) and in TIMS one item [(Item T1b No 148) with one item (no 94) having been excluded from the international TIMSS analysis] were removed from the calibration analyses due to the misfitting of these items to the Rasch model (see Appendices, C and D).

1. 4. 2. Case Estimates

The other way of investigating the fit of the Rasch scale to data is to

examine the estimates for each case. The case estimates give the performance level of each student on the total scale. In order to identify whether the cases fit the scale or not, it is important to examine the case OUTFIT mean square statistic (OUTFIT MNSQ) which measures the consistency of the fit of the persons to the student characteristic curve for each student, with special consideration given to extreme items. In this study, the general guideline used for interpreting t as a sign of misfit is if t> 5 (Wright and Stone, 1979, 169). Thus, if the OUTFIT MNSQ value for a person has a _embed Equation ___ greater than 5, that person does not fit the scale and is deleted from the analysis. In this analysis no person was deleted, because the _embed Equation ___ for all cases was less than 5. However, students with a zero score or with a perfect score were automatically excluded from the calibration procedure.

1. 4. 3. Conclusion

In summary, the results of the infit mean square indices, revealed that 68 out of 70 items for FIMS, 70 out of 72 items for SIMS and 156 out of 157 items for TIMS data sets fitted the Rasch model. In addition, the evidence indicated that for all cases, the responses of the students sampled fitted the Rasch model, except for those students who had perfect or zero scores.

1. 5. Equating of mathematics achievement over time

Equating of the mathematics tests require common items between occasions, that is between FIMS, SIMS and TIMS. Wright and Stone (1979)

have recommended that 10 to 20 (17 to 34 per cent of the items in each test) items should be employed for equating two different test forms consisting of 60 items each. Meanwhile, Hambleton et al., (1991) suggested approximately between 20 and 25 per cent of the number of the items in the tests should be common. However, Smith and Kramer (1992) have argued that as few as a single item is required.

In this study, the number of common items in the mathematics test for FIMS and SIMS data sets were 65. For the mathematics tests the common items formed approximately 93 per cent of the items for FIMS, and 90 per cent for SIMS. Thus, the common items in the mathematics test for these two occasions were well above the percentage ranges proposed by Wright and Stone (1979) and Hambleton et al. (1991).

There were also some items which were common for FIMS, SIMS and TIMS data sets. Garden and Orpwood (1996, 2-2) reported that achievement in TIMSS was intended to be linked with the results of the two earlier IEA studies. Thus, in the TIMS data set there were nine items which were common for the three occasions. Therefore, it was possible to claim that there were sufficient numbers of common items to equate the mathematics test on the three occasions.

Rasch model equating procedures were employed for equating the three data sets. Rentz and Bashaw (1975), Beard and Pettie (1979), Sontag (1984) and Wright (1995) have argued that Rasch model equating procedures are better than other procedures for equating achievement

tests. All three types of Rasch model equating procedures, namely concurrent equating, anchor item equating and common item difference equating were used for equating the three data sets.

Concurrent equating was employed for equating the data sets from FIMS and SIMS. In this method, the 65 common items between FIMS and SIMS were combined into one data set. Hence, the analysis was done on a single data file. Only one misfitting item was deleted at a time so as to avoid dropping some items that might eventually prove to be good fitting items. The acceptable infit mean square values were between 0.77 and 1.30 (Adams and Khoo, 1993). The concurrent equating analyses revealed that among the 65 common items 64 items fitted the Rasch model. Therefore, the threshold values of these 64 items were used as anchor values (see Appendix E) in the anchor item equating procedures employed in the scoring of the FIMS and SIMS data sets separately. Among the 64 common items, nine were common to the FIMS, SIMS and TIMS data sets. The threshold values of these nine items generated in this analysis are presented in Table 2 and were used in equating the FIMS data set with TIMS data sets.

The design of TIMS was different from FIMS and SIMS in two ways. In the first place, only one mathematics test was administered in both FIMS and SIMS, however, in the 1994 study the test included mathematics and science items and the study was named TIMSS (Third International Mathematics and Science Study). The other difference was that in the first two international studies, the test was designed as one booklet.

Every participant used the same test booklet. Whereas in TIMSS, a rotated test design was used. The test was designed in eight booklets. Garden and Orpwood (1996, 2-16) explained the arrangement of the test in eight booklets as follows.

This design called for items to be grouped into "clusters", which were distributed (or "rotated") through the test booklets so as to obtain eight booklets of approximately equal difficulty and equivalent content coverage. Some items (the core cluster) appeared in all booklets, some (the focus cluster) in three or four booklets, some (the free-response clusters) in two booklets, and the remainder (the breadth clusters) in one booklet only. In addition, each booklet was designed to contain approximately equal numbers of mathematics and science items.

All in all there were 286 unique items that were distributed across eight booklets for Population 2 (Adams and Gonzalez, 1996, 3-2).

In order to investigate the level of mathematics achievement in TIMS, it is necessary to find ways and means for equating these eight booklets. Furthermore, in order to employ any kind of test equating procedure there must be common items between the different booklets. Garden and Orpwood (1996) reported that the core cluster items (six items for mathematics) were common to all booklets. In addition, the focus cluster and free-response clusters were common to some booklets. Thus, it was possible to equate these eight booklets and report the achievement level in TIMS on a common scale.

Hence, from among the Rasch model test equating procedures concurrent

equating was chosen for equating these eight booklets. The purposes of the test equating in TIMS was to investigate the mathematics achievement level of Australian students in TIMS and to compare the result with FIMS and SIMS data sets.

Consequently, concurrent equating procedures were employed for the TIMS data set. Appendix D shows the infit mean square values of the first and the last concurrent equating runs. The result of the Rasch analysis indicated that only one item was deleted from the analysis. The item which was deleted from the analysis was Item T1b (No 148) which was below the critical value of 0.77. All other items fitted well the Rasch model.

Table 2:-Descriptive statistics of the common item difference equating procedure employed in FIMS and TIMS

Out of 157 items, 156 of the TIMS test items fitted well the Rasch model. From the output of the concurrent equating, it was possible to obtain the threshold values of the nine common items in TIMS. These threshold values are shown in Table 2.

The next step involved the equating of the FIMS data set with the TIMS data set using the common item difference equating procedure. In this method the threshold values of the FIMS test generated by the QUEST computer program (Adams and Khoo, 1993) for each state are first subtracted from threshold values of the TIMS test. Then the differences

are summed and divided by the number of anchor test items to obtain a mean difference between FIMS and TIMS for each state (see Table 2). The interesting point to be mentioned is that the difference in threshold values between the two occasions in each of the five states was generally similar. The difference between the state with the highest mean threshold difference and the lowest mean score difference was only 0.18. The highest mean threshold estimate value was registered in WA (1.14) while the lowest was in NSW and VIC, the mean difference score for both states was 0.96 (see Table 2). This result revealed that the common items in the two tests behaved similarly in all the five states.

The grand mean difference was calculated by adding the five states mean difference threshold estimates and dividing them by five. The resulting mean difference across states was 1.03. The grand mean of the differences (1.03) is called the equating constant. The equating constant is subsequently employed in the calculation of the TIMS scores on the FIMS scale. That is the equating constant was subtracted from the Rasch estimated mean score on the TIMS for each state to obtain the adjusted mean value of TIMS for each state. A comparison of achievement over time in the five Australian states using the weighted Rasch estimated scores of 1964, 1978 and 1994 for each state are discussed in the next section.

1. 6. Comparisons of Achievement over Time

The comparisons of the performance of students on the mathematics test

for the three occasions were undertaken for two different subgroups namely: (a) 13-year-old students in government schools, who participated in the FIMS and SIMS studies; and (b) Year 8 government school students who participated in the FIMS and TIMS studies. All SIMS students were 13-year-old students. Meanwhile, some of the FIMS students were 13-year-old students, while others were younger and/or older students who were in Year 8. Therefore, for comparison purposes the FIMS students were divided into two groups, namely: (a) FIMSA - involving all 13-year-old students, and (b) FIMSB - including all Year 8 students. Thus, FIMSA students' results could be compared with SIMS students in the government schools of five states, because all students were 13-year-olds. In the TIMS analyses a decision was made to include only Year 8 students, because they were the only group of students who were common to all participating states. Thus TIMS Year 8 government school students from the five states involved could be compared with FIMSB students, because in both groups the students were at the same year level.

1. 6. 1. Comparison between students in mathematics achievement over time

The first part of this section addresses the comparisons between FIMSA and SIMS, while the second part discusses the comparison between FIMSB and TIMS.

Table 3:- Comparisons between FIMS and SIMS 13-year-old Government School Students

1. 6. 1. 1. Comparison between 13-year-old students' mathematics achievement over time

In this section the achievement of 13-year-old Australian students who participated in FIMS and SIMS are compared. Table 3 presents the results of the analyses of the comparison between the two occasions. The first and second panels of the table show the participating states, the estimated case means of the 13-year-old students, the standard deviations and the standard error values, the sample sizes, design effects and effective sample sizes for FIMS and SIMS respectively. While the third panel presents the estimated mean differences between the two groups, the effect sizes and t-values of the differences and the significant levels.

Figure 1:- Comparison of Achievement in Mathematics between 1964 and 1978 in Australia

1.6.1.1.1. State A

When the 1964, 13-year-old State A students estimated mean score is compared with the 1978 same age group students in the same state, the mean score of the 1964 students (458) was higher than that of their 1978 peers (442). The difference was 16 centilogits (see Figure 1 and Table 3). The differences in standard deviation and standard error

values for the two groups were slight, while the design effect was larger in 1964 than in 1978. The effect size was trivial (0.16) and the t-value was 1.39. The estimated mean difference indicated that the mathematics achievement of 13-year-old students in State A had declined over time. However, the effect size and t-values showed that the difference was not practically or statistically significant. Hence, it is possible to conclude that there was no significant decline in mathematics achievement in State A at the 13-year-old student level.

1.6.1.1.2. State B

Table 3 and Figure 1 indicate that when the 13-year-old State B students' who participated in the 1964 First International Mathematics Study, estimated mean score is compared with the mean score of the 1978 same age group students who participated in the Second International Mathematics Study, the 1964 students (483) were found to be higher achievers than their 1978 peers (472). However, the difference was slight, 11 centilogits (see Figure 1 and Table 3). The differences in the standard deviation values for the two groups were also slight, while the standard error and design effect were large. Both were larger in 1964 than in 1978. The effect size was trivial (0.11) and the t-value was 0.70. The estimated mean difference in scores indicated that the mathematics achievement of 13-year-old students in State B declined only slightly over time, since, the effect size and t-values showed that the difference was not practically or statistically significant. Therefore, there was no significant decline in mathematics achievement in State B at the 13-year-old student level.

1.6.1.1.3. State C

The next state that was considered in the comparison between 1964 and 1978 was State C. The estimated mean score of the 1964 13-year-old State C students was 423, meanwhile, the same age group students in 1978 scored 428 (see Table 3 and Figure 1). The mean score difference between the two group was five centilogits in favour of the 1978 students. This indicated that unlike State A and State B, in State C the achievement level of 13-year-old students increased over time (see Figure 1 and Table 3). The differences in standard deviation values for the two groups was slight, while the standard errors and the design effects were larger in 1964 than in 1978. The effect size was trivial (0.05) and the t-value was 0.14. The estimated mean difference indicated that the mathematics achievement of 13-year-old students in State C had improved very slightly over time. However, the effect size and t-value showed that the difference was neither practically nor statistically significant. Hence, it was possible to conclude that there was no significant improvement in mathematics achievement in State C at the 13-year-old student level between 1964 and 1978.

1.6.1.1.4. State D

State D was one of the five Australian states that participated in both the 1964 and 1978 international mathematics studies. The estimated mean score value of the 1964 13-year-old State D students was compared with the 1978 same age group government school students in that state. The mean score difference between students in the two studies was 36

centilogits and the difference was in favour of the 1964 students (see Figure 1 and Table 3). This showed that the achievement of the 1978 students was noticeably lower than that of the 1964 students. In other words, achievement had declined from 1964 to 1978 in the State D government schools. The differences in standard deviation and standard error values for the two groups were slight, while the design effect was larger in 1978 than in 1964. The effect size was small (0.37) and the t-value was 3.12. The estimated mean difference indicated that the mathematics achievement of 13-year-old students in State D government schools had declined over time. In addition, the effect size and t-values also showed that the difference was both practically and statistically significant at the 0.01 level. Hence, it would seem possible to conclude that there was a significant decline in mathematics achievement in State D in government schools at the 13-year-old students level from 1964 to 1978, and that the decline in mathematics achievement represented more than one year's learning of mathematics in the lower secondary schools of Australia.

1.6.1.1.5. State E

State E was the last state for the comparison of performance between the 1964 and 1978 13-year-old students who participated in FIMS and SIMS respectively. When the estimated mean score value of the 1964, 13-year-old State E students was compared with the 1978 same age group government school students in the same state there was no difference in

their mean scores. The mean scores of both groups was 444 (see Figure 1 and Table 3). There was no difference between the achievement of 13-year-old students in State E government schools between 1964 and 1978. The differences in standard deviation and standard error values for the two groups were slight, while the design effect was larger in 1964 than in 1978. The effect size and the t-value were both 0.00. Hence, it would seem possible to conclude that there was no difference in mathematics achievement in State E government schools at the 13-year-old students level over the 14-year period.

The above comparisons were for 13-year-old students in the five states between 1964 and 1978. Among the five states, even if it was not statistically significant, it was only in State C, that achievement over time improved slightly. However, there was no difference between the two occasions in State E. Moreover, in the remaining three states, that is in State A, State B and State D, achievement over time declined, but the decline was significant at the 0.01 level only for State D.

1.6.1.1.6. Australia

The results addressed in Sections 1.6.1.1.1 to 1.6.1.1.5, led to the comparison of the overall Australian 13-year-old students between the two occasions. The estimated mean score difference between the two occasions was 19 centilogits and the difference was in favour of the 1964 13-year-old Australian students. This revealed that the mathematics achievement of Australian students declined from 1964 to

1978. The differences in standard deviation and standard error values for the two groups were small, while the design effect was slightly larger in 1964 than in 1978. The effect size was not inconsiderable (0.19) and the t-value was 2.91. Hence, the mean difference was statistically significant at the 0.01 level (see Table 2 and Figures 1 and 3). Moreover, in Australia the mathematics achievement level of the 13-year-old students declined over time, between 1964 and 1978, to an extent that represented approximately two-thirds of a year of mathematics learning.

In conclusion, the comparisons of the mathematics achievement of the 13-year-old students between 1964 and 1978 in the five Australian states and overall in Australia revealed that in three states and in Australia overall achievement had declined over time. Statistically significant declines were recorded only for State D and for Australia overall. There was no difference between the two occasion for State E students. While, an improvement was recorded for State C, the increase was slight, and it was not statistically significant. The next section presents the comparison of mathematics achievement at the Year 8 level between 1964 and 1994 for students in the government schools of the five states.

1. 6. 1. 2. Comparison between Year 8 students mathematics achievement over time

In Section 1.6.1.1. the mathematics achievement of 13-year-old students in the five Australia states and overall Australia was compared between

FIMS and SIMS. In this section the achievement level of the Year 8 students between 1964 and 1994 are compared. The results of the comparisons of students by state are presented in Table 4 and Figure 2.

1.6.1.2.1. State A

The first state which was selected for comparison was State A. The estimated mean score difference between the two occasions at the Year 8 level in State A was two centilogits, the difference was in favour of the 1964 students (see Figure 2 and Table 4). This result indicated that the mathematics achievement at the Year 8 level had declined very slightly between 1964 and 1994 in State A government schools. The effect sizes and t-values were too small to be considered, and this decline in achievement over time in State A schools at the Year 8 level was not found to be statistically significant.

1.6.1.2.2. State B

State B was the next state that participated in the three international mathematics studies. When the estimated mean scores of the FIMSB and TIMS groups were compared, the 1964 students mean score was noticeably higher than that of the 1994 students. This revealed that mathematics achievement over time had declined in State B schools at the Year 8 level. The standard deviation, standard error and the design effect were larger in 1994 than in 1964. The effect size (0.83) and t-value (4.22) were large. Hence, the decline in mathematics achievement at the Year 8 level between 1964 and

Table 4:- Comparisons between FIMS and TIMS Year 8 Government School Student

were larger in 1994 than in 1964. The effect size (0.83) and t-value (4.22) were large. Hence, the decline in mathematics achievement at the Year 8 level between 1964 and 1994 was practically and statistically significant at the 0.01 level. Moreover, it should be noted that the decline in mathematics achievement in this state represented more than two years of learning.

1.6.1.2.3. State C

The mathematics achievement difference between 1964 and 1994 in State C schools at the Year 8 level was small. The mean difference was 27 centilogits in favour of the 1964 Year 8 students. The result indicated that the Year 8 students' mathematics achievement in State C had declined over time. The standard deviation, standard error and the design effect were larger in 1994 than in 1964. The effect size (0.28) was small, but the t-value (1.46) was not significant. Thus, the t-value indicated that the decline in mathematics achievement between the 1964 and 1994 Year 8 State C students was

Figure 2:- Comparison of Achievement in Mathematics between 1964 and 1994 in Australia

not statistically significant. Hence, it would seem possible to

conclude that there was no statistically significant decline in achievement between 1964 and 1994 in State C at the Year 8 level. However, a substantial decline would seem to have occurred since the effect size (0.28) represented approximately three quarters of a year of mathematics learning.

1.6.1.2.4. State D

The next comparison was between the State D students, and the mean score difference between 1964 and 1994 Year 8 students was 61 centilogits. The difference was in favour of the 1964 students. This indicated that the mathematics achievement level of Year 8 State D school students had declined by more than a year and a half of mathematics learning over the last 30 years. This difference was marked. The effect size was medium (0.59) and the t-value was also large (3.47). Hence, the difference was statistically significant at the 0.01 level. The standard deviation and standard error were larger in 1994 than in 1964. However, the design effect was larger in 1964 than in 1994. Thus, in State D government schools the mathematics achievement level of Year 8 students had declined substantially over the last three decades.

1.6.1.2.5. State E

The last state for comparison between the 1964 and 1994 Year 8 students who participated in FIMS and TIMS respectively was State E. When the estimated mean score value of the 1964 State E Year 8 students was compared with that of the 1994 students, the mean score of the 1994

students was higher than that of the 1964 students (see Figure 2 and Table 3). The difference was 13 centilogits. The finding indicated that in State E schools the mathematics achievement level of Year 8 students had improved over the last three decades. The standard deviation, standard error value and design effect were markedly larger in 1994 than in 1964. The effect size (0.14) and the t-values (0.74) were too small to be considered significant. Hence, it would seem possible to conclude that while there was no statistically significant difference in mathematics achievement in State E schools at the Year 8 level between 1964 and 1994, some signs of improvement had occurred in marked contrast to the other four states, and that the gain was estimated to be approximately half a year of mathematics learning.

The comparisons in the mathematics achievement level of Year 8 students between 1964 and 1994 in State A, State B, State C, State D and State E revealed that only State E showed improvement in mathematics achievement over the last 30 years. However, the improvement was not found to be statistically significant. Moreover in State B, State C and State D the achievement of Year 8 students had declined over the past three decades. A significant decline was recorded in both State B and State D, although the decline in State C was not statistically significant. The next comparison is between Australian Year 8 students on the two occasions.

1.6.1.2.6. Australia

The estimated mean score of the 1964 Australian Year 8 students was 451, while, it was 426 in 1994. The difference was 31 centilogits in favour of the 1964 students (see Table 4, Figures 2 and 3). This difference revealed that the mathematics achievement level of Australian Year 8 students had declined over the 30 year period. The standard deviation, standard error and the design effect were larger in 1994 than in 1964. The effect size was 0.29 and the t-value was 2.16. The effect size and t-value indicated that the decline in mathematics achievement between the 1964 and 1994 Year 8 Australian students was marginally significant at the 0.05 level, and the size of the decline was a little less than a year of mathematics learning.

Table 5:- Comparisons of standard deviation values between FIMS and TIMS

1. 6. 1. 3. Comparison of Standard Deviation

Table 3 shows the standard deviation values for each state in TIMS and FIMS. There would appear to be a large increase in the spread of scores as measured on the scale of mathematics achievement between 1964 and 1994. This increase may be a consequence of less accurate measurement since in 1964 students answered 70 test items while in 1994 the students answered between 33 and 41 items. However, it would seem more probable that there was greater variability in students' mathematical achievement in 1994 compared with 1964 at the Year 8 level as a consequence of changed teaching and learning practices. This issue

warrants further investigation.

1. 7. Conclusion

In order to investigate the mathematics achievement level of lower secondary school Australian students over time, three different data sets, namely from the FIMS, SIMS and TIMS studies were analysed. From the three data sets two groups of students were compared. The first comparison was between 13-year-old government school students in five states who participated in FIMS and SIMS. The result of the comparison revealed that only State C showed improvement in mathematics achievement over the 14-year period. However, the improvement was not statistically significant. Furthermore, no achievement difference was found in State E between 1964 and 1978. Meanwhile mathematics achievement showed a decline in State A, State B and State D. Among the three states a significant decline was found only in State D. When the overall Australian students' performance was compared between 1964 and 1978, the mathematics achievement level of the 13-year-old students declined over the 14-year period (see Figure 3).

Figure 3:- Comparison of Achievement in Mathematics between 1964, 1978 and 1994 in Australia

The second comparison was the mathematics achievement level of Year 8 government school students between 1964 and 1994 in the government schools of five states. The findings indicated that State A and State

E have improved in mathematics achievement over the last 30 years, however, the improvement was not found to be statistically significant. Whereas, in State B, State C and State D the achievement of Year 8 students declined over the last three decades. Significant declines were recorded in State B and State D. However, the decline in State C was not significant. When Australian Year 8 government school students who participated in FIMS and TIMS were compared the decline in mathematics achievement level was found to be marginally significant over the last 30-year period (see Figure 3), but of the order of a little less than a year of mathematics learning in Australian lower secondary schools.

The findings in both comparisons revealed that the mathematics achievement level of Australian students at the lower secondary school level have declined over the last three decades (see Figure 3). The findings also indicate that there is a need to investigate differences in conditions of learning. Carroll's (1963) model of school learning has guided IEA studies and could guide this investigation. Carroll (1963) has identified five factors that influence school learning. These factors are divided into two levels, namely student and school level factors. The student level factors in Carroll's model are aptitude (home background), ability and perseverance (motivation, attitudes). While the school level factors are time for learning (including homework time for mathematics) and quality of instruction. The investigation demands the use of both:

(1) multivariate analysis, and

(2) multilevel analysis.

Thus, it is necessary to conduct further research to identify the reasons and to recommend solutions for the problems.

Acknowledgment

The first author was sponsored by The Flinders University of South Australia Overseas Postgraduate Research Scholarship and the Flinders University Research Scholarship.

Reference

Adams, R. J. & Khoo, S.T. (1993). Quest- The interactive test analysis system. Hawthorn, Victoria: ACER.

Adams, R. J. & Gonzalez, E. J. (1996). The TIMSS test design. In M O Martin & D L Kelly (eds.), Third International Mathematics and Science Study Technical Report vol. 1, Boston: IEA, pp. 3-1 - 3-26.

Anderson, L. W. (1994). Attitude Measures. In T. Husén (eds.), The International Encyclopedia of Education, vol. 1, (second ed.), Pergamon, pp. 380-390.

Beard, J. G. & Pettie, A. L. (1979). A comparison of Linear and Rasch Equating results for basic skills assessment Tests. Florida: Florida state university: ERIC.

Brick, J. M., Broene, P., James, P. & Severynse, J. (1997). A user's guide to WesVarPC. (Version 2.11). Boulevard, MA: Westat, Inc.

Byrne, B. M. (1989). A primer of LISLEL basic applications and programming for confirmatory factor analytic models. New York:

Springer-Verlag.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112 (1), 155-159.

Garden , R. A. & Orpwood, G. (1996). Development of the TIMSS achievement tests. In M 0. Martin & D L Kelly (eds.), Third International Mathematics and Science Study Technical Report Volume 1: Design and Development, Boston: IEA, pp. 2-1 to 2-19.

Hambleton, R. K.& Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. Journal of educational measurement, 14 (2), 75-96.

Hambleton, R. K., Zaal, J. N.& Pieters, J. P. M. (1991). Computerized adaptive testing: theory, applications, and standards. In R K Hambleton & J N Zaal (eds.), Advances in Educational and Psychological Testing, Boston, Mass.: Kluwer AcademicPublishers, pp. 341-366.

Keeves, J. P. (1992). The design and conduct of the second science study. In J P Keeves (eds.), The IEA Study of Science III: Changes in Science Education and Achievement: 1970 to 1984, Oxford: Pergamon, pp. 42-67.

Kim, J & Mueller, C. W. (1978a). Introduction to Factor Analysis What It Is and How to Do It.. London: Sage Publications.

Kim, J & Mueller, C. W. (1978b). Factor Analysis Statistical Methods and Practical Issues. London: Sage Publications.

Kolen, M. J. & Whitney, D. R. (1981). Comparison of four procedures for equating the tests of general educational development. Paper presented at the annual meeting of thee American Educational Research Association. Los Angeles, California.

Lokan, J., Ford, P. & Greenwood, L. (1996). Maths & Science On the Line: Australian Junior Secondary Students' Performance in the Third International Mathematics and Science Study. Camberwell: ACER.

Long, J. S. (1983). Confirmatory Factor Analysis: A preface to LISREL. Beverly Hills: Sage Publications.

Moss, J. D. (1982). Towards Equality: Progress by Grls in Mathematics in Australian Secondary Schools. Hawthorn, Victoria: ACER.

Norusis, M. J. & SPSS Inc (1993). SPSS for windows: Base system user's guide: Release 6.0. Chicago: SPSS Inc.

Peaker, G. F. (1969). How should national part scores be weighted? International Review of Education, 15, 229-237.

Rentz, R. R. & Bashaw, W. L. (1975). Equating Reading tests with the Rasch model, Vol. I Final Report. Athens, Georgia: University of Georgia: Educational Research Laboratory, College of Education.

Rosier, M. J. (1980). Changes in Secondary School Mathematics in Australia. Hawthorn, Victoria: ACER.

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

Sontag, L. M. (1984). Vertical equating methods: A comparative study of their efficacy. DAI, 45-03B, page 1000.

Spearritt, D. (1994). Factor Analysis. In T Husén & T.N Postlethwaite (eds.), The International Encyclopedia of Education, (second ed.), vol.4. Oxford: Pergamon, pp. 2230-2241.

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

Wright, B. D., and Stone, M. H. (1979). Best Test Design: Rasch Measurement. Chicago: Mesa Press.

Willett, J. B. (1997). Change, Measurement of. In J P Keeves (ed.), Educational Research, Methodology, and Measurement: An International Handbook, (second ed.), Oxford: Pergamon, pp. 327-334.

Appendix A:- Errors of Estimations and Scaling

In the present study, sources of errors of estimation and scaling that are related to the calculation of gains and losses in mathematics achievement are associated with the sampling design, the fitting of individual items to a scale based on the Rasch model (calibration) and the use of the equating constant based on the FIMS and TIMS data sets.

(1). The error associated with the sampling design for each data set was generated using WesVarPC computer program (Brick, et al., 1997).

(2). The error associated with the use of the mean value of the equating constant arises from the equating using the nine common items in the FIMS and TIMS mathematics tests in the five state samples.

The error associated with the equating constant was estimated to be 0.104. The items employed for anchoring in the common item difference equating procedure are not a random sample of items but a fixed sample of specifically chosen items. Under these circumstances the error of the grand mean is given by _embed Equation ___ where n is the number of items in the sample for each state.

Standard error of equating constant = _embed Equation ___

(3). For individual students the QUEST computer program (Adams and Khoo, 1993) provided a value for the magnitude of the measurement error associated with the estimation of student performance. This estimate for TIMS was about 35 scale units. In order to calculate the error arising from calibration the following formula was used:

The QUEST computer program (Adams and Khoo, 1993) by default does not process cases with perfect and zero scores, because both groups do not provide information for the calibration of the scale. Hence, in order to include those cases with perfect and zero scores in the calculation of the mean and standard deviation of the mathematics achievement test scores for each student sampled, the values of the perfect and zero scores were calculated by extrapolation from the logit table produced by the QUEST computer program. Table A shows the procedures employed to estimate the scores of cases with a perfect and zero score. The calculation of the scores of the FIMS students who had perfect and zero scores has been used here as an example. Table A1 shows the procedures employed to estimate the scores of cases with a perfect score. The first column indicates the top three raw scores (69, 68 and 67) excluding the highest possible raw score (70). The second column indicates the logit values obtained from the logit table generated by the QUEST computer program (Adams and Khoo, 1993). This column provides the Rasch scores corresponding to the top three possible raw scores in the test excluding the maximum score. D1 gives the successive differences between the top three logit values. It was assumed that

compared to the highest logit value, the perfect score was likely to be greater than a value equal to the difference between the top two scores and the difference between consecutive differences of the top three scores. Therefore, the following calculation was employed to estimate the perfect score. In order to get the first entry (0.73) in column D1 the second highest logit value (4.17) was subtracted from the first highest logit value (4.90). The same procedure was applied to obtain the second entry (0.44) in which, the third highest value (3.73) was subtracted from the second highest value (4.17). The difference between the two entries in column D1, that is the difference between 0.73 and 0.44, namely 0.29, was entered in column D2. Therefore, the estimated Rasch score for the maximum raw score 70 is assigned in the column Perfect Score in Table A1. The score (5.92) was estimated by adding the highest logit value (4.90) for a score of 69 and the first entry in column D1 (0.73) and the entry in column D2 (0.29).

For the estimation of the zero score it was assumed that compared to the lowest logit value, the zero score would most likely be less than the logit value for a score of one, by a value equal to the difference between the bottom two scores and the difference between consecutive differences of the bottom three scores. Hence, the same procedure was employed for the estimation of the zero score. Table A2 shows the estimation of the zero scores. Thus estimation procedure employed for zero scores was similar to the one applied for the perfect score estimation. However, the subtractions for the estimation of zero score were from the bottom. Thus, to obtain the first entry (-0.73) in column

D1 the second lowest value (-3.97) was subtracted from the first lowest value (-4.70). In order to obtain the second entry (-0.44) in column D1 the third lowest value (-3.53) was subtracted from the second lowest value (-3.97). Moreover, in order to obtain the entry in column D2 the second lowest value of D1 was subtracted from the lowest value of D2. Therefore, the estimated Rasch score for the minimum raw score of zero is assigned in the column Zero Score in Table 2b. The score (-5.72) was estimated by adding the lowest logit value (-4.70) for a score of one and the lowest entry in column D1 (-0.73) and the entry in column D2 (-0.29).

Appendix C:- Infit mean square values (INFIT MNSQ) for Mathematics test items all students in FIMS and SIMS using Anchor Item Equating Procedure

	F I M S		S I M S	
Item No	Before Deletion	After Deletion	Before Deletion	After Deletion
item 1	0.97	0.97	0.96	0.96
item 2	0.81	0.81	0.83	0.83
item 3	0.93	0.93	1.06	1.06
item 4	0.88	0.88	0.86	0.86
item 5	0.94	0.94	1.03	1.03

item 6 | 0.93| 0.93| 0.95 | 0.96|

item 7 | 1.02| 1.03| 0.86 | 0.85|

item 8 | 0.93| 0.93| 0.97 | 0.98|

item 9 | 0.83| 0.83| 0.84 | 0.84|

item 10| 1.03| 1.03| 0.90 | 0.90|

item 11| 0.85| 0.85| 0.80 | 0.80|

item 12| 1.09| 1.09| 1.02 | 1.03|c

item 13| 0.74|Deleted| 0.95 | 0.94|

item 14| 1.19| 1.19| 1.11 | 1.12|

item 15| 1.11| 1.10| 1.04 | 1.05|

item 16| 1.08| 1.08| 1.05 | 1.06|d

item 17| 0.94| 0.94| 0.96 | 0.96|

item 18| 0.97| 0.97| 1.03 | 1.04|

item 19| 0.89| 0.90| 0.80 | 0.79|

item 20| 0.90| 0.90| 0.90 | 0.90|

item 21| 1.24| 1.23| 1.39 |Deleted|

item 22| 0.82| 0.82| 0.94 | 0.94|

item 23| 0.94| 0.94| 0.92 | 0.92|

item 24| 0.87| 0.87| 0.85 | 0.85|

item 25| 0.86| 0.86| 0.87 | 0.87|

item 26| 0.94| 0.94| 0.94 | 0.95|c

item 27| 0.97| 0.97| 0.85 | 0.86|

item 28| 1.00| 1.00| 0.96 | 0.96|

item 29| 1.34|Deleted| 0.74 |Deleted|

item 30| 0.83| 0.82| 1.13 | 1.13|

item 31| 0.82| 0.82| 1.09 | 1.09|c

item 32| 0.90| 0.90| 0.89 | 0.90|c

item 33| 0.82| 0.82| 0.87 | 0.87|c

item 34| 0.96| 0.95| 0.89 | 0.89|

item 35| 1.15| 1.15| 1.05 | 1.06|

item 36| 1.01| 1.01| 1.04 | 1.05|

item 37| 1.04| 1.04| 1.00 | 1.00|

item 38| 0.92| 0.92| 0.95 | 0.95|c

item 39| 0.94| 0.94| 0.90 | 0.90|

item 40| 0.92| 0.92| 0.96 | 0.96|

item 41| 1.09| 1.09| 1.11 | 1.12|d

item 42| 0.77| 0.77| 1.09 | 1.08|

item 43| 1.22| 1.21| 1.05 | 1.06|d

item 44| 1.08| 1.07| 1.13 | 1.14|

item 45| 1.18| 1.18| 0.98 | 0.99|d

item 46| 0.97| 0.97| 0.99 | 0.99|

item 47| 0.80| 0.80| 0.92 | 0.93|

item 48| 0.85| 0.85| 0.87 | 0.87|

item 49| 1.10| 1.10| 1.09 | 1.09|

item 50| 0.87| 0.87| 0.90 | 0.90|

Continued..

Appendix C: (Continued)

==========|

| F I M S | S I M S |

=======|===============|===============|

Item |Before |After |Before |After |

No |DeletionDeletionDeletion Deletion

=======|=======|=======|========|=======|

item 51| 0.81| 0.81| 0.80 | 0.80|

item 52| 0.79| 0.79| 0.95 | 0.96|

item 53| 0.87| 0.86| 0.87 | 0.87|

item 54| 0.98| 0.98| 0.92 | 0.92|c

item 55| 0.91| 0.91| 0.92 | 0.93|

item 56| 0.89| 0.89| 0.93 | 0.93|

item 57| 0.87| 0.87| 0.99 | 0.98|

item 58| 0.91| 0.91| 0.92 | 0.93|

item 59| 1.18| 1.18| 0.98 | 0.99|d

item 60| 1.12| 1.12| 1.11 | 1.12|

item 61| 1.13| 1.13| 1.25 | 1.26|

item 62| 0.99| 0.99| 1.03 | 1.04|

item 63| 0.94| 0.94| 1.08 | 1.08|

item 64| 0.94| 0.93| 1.11 | 1.11|

item 65| 1.09| 1.08| 1.15 | 1.16|

item 66| 0.99| 0.98| 1.20 | 1.21|

item 67| 1.01| 1.01| 0.98 | 0.99|c

item 68| 0.92| 0.92| 0.95 | 0.96|

item 69| 0.97| 0.97| 0.98 | 0.98|

item 70| 1.08| 1.08| 1.18 | 1.19|

item 71| © | | 1.05 | 1.06|

item 72| | | 0.80 | 0.80|

==========|=======|======= ========|

Mean | 0.97 | 0.96| 0.98 | 0.98|

SD | 0.12 | 0.11| 0.12 | 0.11|

N | 4320 | 4320 | 5120 | 5120 |

==========

SD = standard deviation

d= Different items were administered for each occasion, therefore, the

items were not anchor items

© = 70 items were administered for FIMS while 72 for SIMS

c= Common items for FIMS, SIMS and TIMS

Appendix D: Infit mean square values for Mathematics test items Year 8

students in TIMS using Concurrent Equating procedure

==========|==========|==========

==========

|Before |After | |Before |After |

|Before |After

Items |Deletion |Deletion| Items |Deletion |Deletion| Items

|Deletion |Deletion

==========|==========|==========

==========

item 1 | 0.81 | 0.81 | item 54 | 0.99 | 0.99 | item 107

| 1.07 | 1.07

item 2 | 1.02 | 1.01 | item 55 | 1.02 | 1.01 | item 108

| 0.82 | 0.82

item 3 | 0.92 | 0.92 | item 56 | 1.15 | 1.14 | item 109

| 0.91 | 0.91

item 4 | 1.00 | 1.00 | item 57 | 0.92 | 0.92 | item 110

| 0.95 | 0.95

item 5 | 1.22 | 1.22 | item 58 | 0.86 | 0.86 | item 111

| 1.01 | 1.01

item 6c | 1.08 | 1.08 | item 59 | 1.12 | 1.12 | item 112

| 0.86 | 0.86

item 7 | 1.00 | 0.99 | item 60 | 0.98 | 0.98 | item 113

| 0.94 | 0.94

item 8 | 1.28 | 1.28 | item 61 | 0.98 | 0.98 | item 114

| 1.20 | 1.20

item 9 | 1.07 | 1.07 | item 62c | 1.01 | 1.01 | item 115

| 1.05 | 1.05

item 10 | 0.82 | 0.81 | item 63 | 1.07 | 1.07 | item 116

| 1.16 | 1.16

item 11 | 0.96 | 0.96 | item 64 | 1.07 | 1.07 | item 117

| 0.95 | 0.95

item 12 | 0.87 | 0.87 | item 65 | 1.03 | 1.03 | item 118

| 1.01 | 1.01

item 13 | 0.99 | 0.99 | item 66 | 1.15 | 1.15 | item 119

| 0.91 | 0.91

item 14 | 0.91 | 0.91 | item 67 | 1.08 | 1.07 | item 120

| 0.99 | 0.99

item 15 | 0.89 | 0.89 | item 68 | 0.93 | 0.93 | item 121

| 0.86 | 0.86

item 16 | 1.28 | 1.28 | item 69 | 1.05 | 1.05 | item 122

| 1.27 | 1.27

item 17 | 0.97 | 0.97 | item 70c | 1.17 | 1.15 | item 123

| 1.08 | 1.08

item 18 | 1.00 | 1.00 | item 71 | 0.80 | 0.79 | item 124

| 1.12 | 1.12

item 19 | 0.99 | 0.99 | item 72 | 1.16 | 1.15 | item 125

| 1.11 | 1.11

item 20 | 0.97 | 0.97 | item 73 | 0.95 | 0.95 | item 126

| 0.96 | 0.96

item 21 | 0.83 | 0.83 | item 74 | 1.07 | 1.06 | item 127

| 0.94 | 0.94

item 22 | 1.22 | 1.21 | item 75 | 0.83 | 0.83 | item 128

| 0.98 | 0.98

item 23 | 0.94 | 0.94 | item 76 | 1.06 | 1.06 | item

129c| 0.95 | 0.95

item 24 | 1.03 | 1.03 | item 77 | 0.92 | 0.92 | item 130

| 0.80 | 0.80

item 25 | 1.05 | 1.05 | item 78 | 0.97 | 0.97 | item 131

| 0.99 | 0.99

item 26 | 1.03 | 1.03 | item 79 | 1.19 | 1.19 | item 132

| 0.97 | 0.97

item 27 | 1.00 | 1.00 | item 80 | 0.87 | 0.87 | item 133

| 0.95 | 0.95

item 28 | 0.89 | 0.89 | item 81 | 0.94 | 0.94 | item 134

| 1.10 | 1.10

item 29 | 0.88 | 0.88 | item 82 | 1.13 | 1.13 | item 135

| 1.06 | 1.06

item 30 | 1.00 | 1.00 | item 83 | 1.05 | 1.05 | item

136c| 0.96 | 0.96

item 31c | 1.12 | 1.12 | item 84 | 0.88 | 0.88 | item 137

| 0.95 | 0.95

item 32 | 1.11 | 1.11 | item 85 | 0.91 | 0.91 | item 138

| 1.04 | 1.04

item 33 | 0.96 | 0.96 | item 86 | 0.93 | 0.93 | item 139

| 0.84 | 0.84

item 34 | 0.96 | 0.96 | item 87 | 1.10 | 1.10 | item 140

| 0.89 | 0.89

item 35 | 0.85 | 0.85 | item 88 | 1.03 | 1.03 | item 141

| 0.79 | 0.79

item 36 | 1.05 | 1.05 | item 89 | 0.93 | 0.93 | item 142

| 1.03 | 1.03

item 37 | 1.13 | 1.12 | item 90 | 1.08 | 1.08 | item 143

| 0.93 | 0.92

item 38 | 0.99 | 0.99 | item 91 | 0.91 | 0.91 | item 144

| 0.83 | 0.83

item 39c | 1.05 | 1.05 | item 92c | 1.03 | 1.03 | item 145

| 0.81 | 0.80

item 40 | 0.95 | 0.95 | item 93 | 1.00 | 1.00 | item 146

| 0.89 | 0.88

item 41 | 0.97 | 0.96 | item 94 Excluded Item | item 147

| 0.96 | 1.04

item 42c | 0.99 | 0.99 | item 95 | 0.92 | 0.92 | item 148

| 0.75 |Deleted

item 43 | 1.07 | 1.07 | item 96 | 1.08 | 1.08 | item 149

| 0.85 | 0.85

item 44 | 0.89 | 0.89 | item 97 | 0.92 | 0.92 | item 150

| 1.02 | 1.02

item 45 | 0.95 | 0.95 | item 98 | 1.01 | 1.01 | item 151

| 1.01 | 1.01

item 46 | 0.89 | 0.89 | item 99 | 1.27 | 1.27 | item 152

| 1.00 | 0.99

item 47 | 1.00 | 1.00 | item 100 | 1.12 | 1.12 | item 153

| 1.09 | 1.08

item 48 | 0.95 | 0.95 | item 101 | 1.14 | 1.14 | item 154

| 0.91 | 0.90

item 49 | 1.23 | 1.22 | item 102 | 0.91 | 0.91 | item 155

| 0.83 | 0.83

item 50 | 0.99 | 0.98 | item 103 | 0.78 | 0.78 | item 156

| 1.06 | 1.06

item 51 | 1.29 | 1.28 | item 104 | 0.96 | 0.96 | item 157

| 1.02 | 1.02

item 52 | 1.00 | 0.99 | item 105 | 0.95 | 0.95 | item 158

| 0.92 | 0.92

item 53 | 0.95 | 0.94 | item 106 | 1.02 | 1.02 |