Investigating whether aberrant response behaviour in classroom maths tests is a stable characteristic of students

Panayiotis Panayides [a] and Peter Tymms [b]

[a] *Lyceum of Polemidia,Limassol, Cyprus,* [b] *Durham University, Durham, UK*

Corresponding Author:
Panayiotis Panayides
Nikou Kavadia 1, K. Polemidia
Limassol 4152, Cyprus
Email: p.panayides@cytanet.com.cy

Panayiotis Panayides holds a BSc in Statistics with Mathematics (Queen Mary College, University of London), an MSc in Educational Testing (Middlesex University, UK) and a PHD in Educational Measurement (University of Durham, UK). He is currently an assistant headmaster and head of the Mathematics department at the Lyceum of Polemidia, Limassol, Cyprus. His research interests include educational and psychological measurement and research into mathematics education.

Professor Peter Tymms PhD is Head of Department in the School of Education and was, until recently Director of CEM (Centre for Evaluation and Monitoring) at Durham University. His main research interests include monitoring, assessment, interventions and research methodology generally. He set up the PIPS (Performance Indicators in Primary Schools Project) which runs in thousands of schools around the world. Work in these and related areas have produced more than a hundred publications and reports. Peter Tymms is an adviser to the German National Educational Panel Study, is on the Expert Board of the European Science Foundation and a member of the Academy of Social Science.

Investigating whether aberrant response behaviour in classroom maths tests is a stable characteristic of students

Abstract

This study investigated whether aberrant response behaviour is a stable characteristic of high school students taking classroom maths tests as has been implied in the literature.

For the purposes of the study two maths tests were administered; the first to 25 classes (635 students) and the second to 18 out of the original 25 classes (445 students). The tests contained multistep mathematical problems with partial credit awarding for partially correct answers, together with some multiple-choice items. The Rasch Partial Credit Model was used for the analyses and the infit and outfit mean square statistics with six different cut-off scores were used to identify students with aberrant response behaviour (misfitting students).

Six Chi-square tests were then performed, one for each cut-off score, leading to a very clear conclusion: Contrary to expectations the same students do not misfit in the two tests administered; aberrance does not seem to be a stable characteristic of students.

Explanations for aberrant responses such as carelessness, plodding or guessing need to be reconsidered. They may have validity for particular test situations but this has yet to be demonstrated and thus investigation calls them into question.

Keywords: Aberrance, classroom maths tests, Rasch

Introduction

Classroom achievement tests are made from a set of items administered to pupils through which the teacher evaluates how effectively his or her students have learned what has been taught. They are assessment tools that help teachers in evaluating students' overall achievement and growth in a content domain, assigning grades to students, improving their own teaching methods, diagnosing students' strengths and weaknesses, encouraging good study habits, planning review materials, determining the pace of instruction in the classroom and reporting achievement to parents.

Popham (2000) places emphasis on the contribution of tests in promoting more effective teaching and argues that classroom tests if properly conceptualized, with instruction in mind, are more useful than commercially made tests mainly because of the clarity associated with what is being measured. For this reason, it is common practice for teachers to use tests that they have prepared themselves much more frequently than any other types of test to monitor what has been previously learned.

> "Classroom tests, despite some of their limitations, will never be replaced because they (a) tend to be more relevant, (b) can be tailored to fit a teacher's particular instructional objectives, and (c) can be adapted better to fit the needs and abilities of the students than can commercially published tests." (Mehrens and Lehmann 1991, 79).

It is therefore common practice for classroom maths teachers, at least in high schools in Cyprus, to regularly administer maths tests to their students during each term and particularly at the end of each learning unit. This way the teachers can assess how well their students have learned the unit's material before proceeding to the next.

Aberrant response behavior

An aberrant response pattern is one that is improbable, given either that an IRT model fitted the data or given the item response patterns of other persons in the group. In the context of Rasch measurement a person's aberrant response pattern in a test is one that is improbable, given that the data fit the Rasch model.

Many authors have described several types of individuals whose response patterns do not fit the typical pattern. These types include 'sleepers' who get bored as they proceed in the test and do poorly on the last items (Linacre and Wright 1994; Molenaar and Hoijtink 1996), 'fumblers' who get confused with the item format and do poorly in the beginning of the test (Bracey and Rudner 1992), 'plodders' who take too much time to answer and never get to the later items (Meijer 1996; Wright 1977), 'guessers' who select answers at random and 'cheaters' (Athanasou and Lamprianou 2002; Rudner 1983). The list can be expanded with people who show extreme creativity in interpreting questions (Karabatsos 2000; Meijer 1996) or with poor language skills (Rudner 1983). Other possible reasons at the item level or the administration process include item multidimensionality, item bias, multiple correct options for an item, disordered pages in a booklet or miskeyed items (Karabatsos 2000).

Possible Factors associated with aberrance

Gender

Frary and Giles (1980) showed that overall females had lower person fit statistics values, indicating lower aberrance for this group.

Mismatch between curriculum and test content

Harnisch and Linn (1981) reported that schools in different parts of Illinois had very different fit indices. They attributed this school effect to the possibility that certain schools may have not covered segments of the content sampled by the test, or that they

may have given less emphasis to some of the content, suggesting that the differences in the index were caused by a mismatch between school curriculum and test content.

Position on the ability/trait scale

Keeves and Masters (1999) expressed concerns about trait range affecting misfit, suggesting that persons in different ability ranges could have different proportions of misfits. However, Li and Olejnik (cited in Curtis 2004, 130) found no correlation between trait estimate and misfit. On the other hand, Petridou and Williams (2007) report that high ability students can manifest more aberrance and this can be attributed to carelessness and silly mistakes.

Test anxiety

Various authors report test anxiety as a possible source of aberrance (Athanasou and Lamprianou 2002; Bracey and Rudner 1992; Harnisch and Linn 1981). Harnisch and Linn (1981) suggest that test anxiety may make normally simple items seem very difficult to some people, and Emons et al. (2003) suggest that test anxiety may result in many errors in the first items of the test, implying that after the first items test anxiety decreases.

Motivation

Lamprianou and Boyle (2004) suggest that examinees with too little motivation may be potentially more likely to produce aberrant response patterns and suggest that the number of unauthorized absences may be considered as an indication of atypical schooling or low motivation. Bracey and Rudner (1992) also suggest atypical schooling as a possible factor associated with aberrance.

Class effect

Petridou and Williams (2007) reported a high class level effect on aberrance. They suggest that non-standard administration practices such as teachers interpreting questions, class 'cheating' (p. 243) by leaving materials related to the test on the classroom walls and instructional effects in terms of topics not being taught by the time of test administration are possible reasons for this significant effect.

Panayides (2009) investigated possible associations of misfit in classroom maths tests with the above-mentioned factors together with item order, Attention Deficit Hyperactivity Disorder (ADHD), maths self-esteem, language competency and study habits, and found none. He attributed his findings to the fact that the test items were mainly multistep mathematical problems with partial credit awarding for partial success instead of the usual dichotomous items found in the literature and to the low status of the tests and the non-stressful administration procedure in the classroom setting.

Smith (1986) and Lamprianou (2005) suggested that an individual with an aberrant response pattern may exhibit such response behaviour in other testing situations too, implying that misfit could be a stable characteristic of individuals. Moreover, many researchers (such as Athanasou and Lamprianou 2002; Karabatsos 2003; Reise and Flannery 1996; Rudner 1983) have argued that aberrant responses may lead to misleading score interpretations and consequently to invalid measurement.

The present study

The study of aberrant scores is important since it has many potential advantages ranging from improving ability estimates (Levine and Drasgow 1988), diagnosing sources of misfit (Linacre and Wright 1994), analyzing group, schooling and instructional differences (Harnisch and Linn, 1981) or diagnosing causes of low test scores (Wright 1977).

Panayides (2009) found no associations with aberrance of several of the factors reported in the literature. On the other hand Smith (1986) and Lamprianou (2005) suggested that an individual with an aberrant response pattern may exhibit such response behaviour in other testing situations too, implying that misfit could be a stable characteristic of individuals. This is also implied by the various types of individuals, suggested in the literature, that exhibit aberrant response behaviour (sleepers, fumblers, plodders and guessers). This contradiction has led the researchers to this study. Its aim is not to investigate possible factors associated with aberrant response behaviour (as in Panayides, 2009) or to develop procedures to identify specific types of aberrance (as in Lamprianou, 2010). Instead the aim of this study is to investigate whether aberrant response behaviour is a stable characteristic of high-school students in classroom maths tests. In simpler words, whether essentially the same students will misfit in administrations of two different classroom maths tests.

Method

Two maths tests (Test 1 and Test 2) were administered to a sample of first form lyceum students in Cyprus (ages 15-16 years). One of the researchers is a high school maths teacher and, for the purpose of convenience, his school was selected together with a further two, randomly selected, from the 10 lyceums in the town of Limassol.

Prior to commencing data collection, a letter was sent to the Director of Secondary Education at the Ministry of Education and Culture (MOEC), seeking permission to administer the mathematics tests to the sample. The letter offered assurance for the safeguarding of the participants anonymity and clarification that written consent would be sought from the heads, teachers, students and their parents for participation. The Director granted permission stipulating no teaching time be lost during the data collection.

The researchers agreed and provided detailed oral and written explanations regarding the purposes of the study and the role of the teachers in the process of data collection to the participating heads of the Maths departments and teachers, and sought and received their written consent. The teachers in turn informed their students about the study prior to administering the tests, and gave them a consent form to complete and another for their parents. In the first school this practice proved very time consuming and consent forms were lost in the process. Consequently, in the remaining two schools only the written consent of the students was sought. However, in compliance with assurances made by the researchers to the MOEC, the students were asked to inform their parents of their partaking in the study and were reminded of their right to withdraw should their parents object to their participation. No parental objections were voiced and no students withdrew.

Test 1

The first test was a 'diagnostic' test, administered towards the end of September, by 13 teachers, to 635 students in 25 classes of the three schools. Of the 635 students, 279 (43.9%) were male and 356 (56.1%) female. Such a test is always administered at the beginning of the year in lyceums in Cyprus to all first form students. Its purpose is for teachers to get an initial evaluation of their students' abilities but at the same time to identify the very weak students. For this reason the test contains items on very basic skills and abilities, those which teachers feel are the most important for students to possess in order to be able to follow the new year's syllabus. Once the very weak students are identified, they are encouraged to take extra lessons in the subject which are offered free of charge by the school.

Test 1 consisted of 27 items carrying from 1 to 5 marks, giving a total score of 50. Three of these items, 2a, 2b and 2c, were multiple choice questions with three options to choose from, carrying one mark each.

One of the researchers, an experienced teacher of mathematics, prepared the test with the help and suggestions for improvements from two other teachers working in the two other schools. Once prepared, the test was sent to all the teachers participating and their comments were sought. Suggestions for the refinement of the test were brought forward, taken into consideration, and the final refined test was prepared. In its final version, the test contained items on algebra (expansions, algaebraic identities, operations with algebraic fractions, factorisations, solutions of linear equations including ones containing algebraic fractions and an item on straight line graphs) and items on geometry (angle properties, types of triangles, angle and side relations in triangles).

The test was administered over one 45-minute teaching period. Each school administered the test simultaneously to all the classes; however the schools chose the date and period of the test independently from one another. To ensure more reliable results a detailed marking scheme was prepared which was thoroughly explained to and discussed with all the teachers so as to leave no questions or ambiguities.

Reliability and validity of Test 1

For the study of the reliability of the test, the student reliability was used. Validation studies included professional input from maths teachers, described above, Principal Components Analysis (PCA) of the standardized residuals, after the Rasch calibrations, as proposed by Linacre (1998) and correlations of the test results with the final maths exam. The latter was done separately for each school, since the three schools had a different final maths exam. Finally, comparisons of the item estimates from two

different calibrations (using two different samples, based on gender) were made to ascertain whether the property of invariance holds. The reason for this split was that the two groups had significant differences (p = 0.044) in their scores. The male mean score was 31.04 **(N = 279, SD = 13.56)** and the female mean scores 33.15 **(N = 356, SD = 12.55).**

Test 2

The second test was on quadratic equations. It consisted of 2 sections. The first section had 12 multiple choice items, carrying 1 mark each and the second section 4 multistep problems carrying 4 marks each. The maximum possible score for this test was 28. It was prepared and administered in exactly the same way as Test 1, with the cooperation of the researchers with teachers from the schools involved. It was administered over a 45-minute teaching period to 18 out of the 25 classes, that is, 445 out of the 635 students who originally took Test1. Overall, out of the total of 445 students, 41.8% were male and 58.2% female. The reason for this smaller sample was that four teachers were not particularly willing to participate further. Test 2 was not administered simultaneously to all classes. Instead, the teachers were free to choose the time when they felt that their students were ready and prepared for it. The researchers did not want to put pressure on the teachers by giving deadlines for the administration of the test. Furthermore, although the curriculum in Cyprus is the same for all the schools, teachers have the freedom to teach it in whichever order they feel is the best for them and their students and the researchers did not want to interfere with that. Finally, a detailed marking scheme was again prepared, thoroughly explained to and discussed with all the teachers involved.

Reliability and validity of Test 2

A content validity questionnaire was administered to eight very experienced mathematics teachers, all with more than 20 years of experience in teaching the subject in public schools. In the questionnaire the experts had to express the degree to which they agreed or disagreed, using a 4-point Likert scale, on statements regarding the clarity of the questions, the adequacy of time to complete the test, the coverage of all the important skills of the specific chapter as described in the syllabus and whether the test contained any items on skills not included in the syllabus.

For the study of the reliability of the test the student reliability was used as in Test 1. For the validation study of the test, PCA of the standardized residuals was used and correlations of the test results with the final maths exam. The latter was again done separately for each school. Also comparisons of the item estimates from two different calibrations (based again on gender, in order to be consistent with the comparisons made in Test 1) were made to investigate whether the property of invariance holds. Finally, comparisons of ability estimates from the two maths tests were made.

Both tests were typical classroom tests because even though they were prepared by a team of classroom maths teachers led by one of the researchers who is himself a maths teacher and at the same time an assessment expert, their objective was to assess each student's ability and to identify possible weaknesses. They were administered by the maths teachers to their classes during a normal 45-minute teaching period. The class teachers marked them, returned them to their students, and provided remedial instruction where they felt it was necessary. Finally, the tests were used as part of the assessment of students in mathematics for the first and second terms of the academic year.

Selection of the Rasch Partial Credit Model (PCM)

The Rasch models are generally easier than other IRT models to work with, to understand and to interpret because they involve fewer parameters; they have fewer parameter estimation problems and give more stable item estimates with smaller samples. They also present ability measures and item calibrations on a common logit scale (Bond and Fox 2001; Wright and Masters 1982) making it easy to see relations between them. Also validity and reliability issues can be addressed by the Rasch models (Smith 2004).

The Rasch PCM (Masters 1982) was selected for use in this study also because by accepting the scores of the examinees as the sufficient statistics for the ability estimates the score order is maintained. Since raw scores are the basis for reporting results throughout the educational system in Cyprus, and especially in classroom tests, this model is consistent with practice. Finally, multistep problems with partial credit awarding for partially correct answers can easily be dealt with by the Rasch PCM.

The model

Masters (1982) introduces the PCM which is given by:

$$\Pi_{xni} = \frac{\exp \sum_{j=0}^{x} (\beta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\beta_n - \delta_{ij})} \qquad x = 0,1,...,m_i$$

where for notational convenience $\sum_{j=0}^{0} (\beta_n - \delta_{ij}) = 0$, Π_{xni} is the probability of person n scoring x on item i, β_n is the person's position on the variable and δ_{ij} are the difficulties of the m_i 'steps' in item i.

Selection of the fit statistics

Infit and outfit mean square statistics have been used to assess the reasonableness of the response patterns in this study because of their exploratory nature (Douglas 1990), their ability to identify a wide range of potential sources of aberrance (Wright 1997), their successful use in assessing the fit to the Rasch models for many years (Curtis 2004; Smith 1990), and their simple computation which makes them more attractive to users. Smith (1990) also reports that they stand up well in comparison with other possibly more precise tests and that gives no practical reason to use anything more complicated. Finally, the two fit statistics are utilized by most of the available software packages for Rasch calibrations (e.g. Quest, Winsteps and Facets) and are familiar to many researchers.

Critical values of the fit statistics

Smith (1996) argues that the aim of the fit statistics is to aid in measurement quality control by identifying those parts of the data that do not meet the Rasch model specifications and could corrupt measurement. Linacre and Wright (1994) explain that fit values noticeably above 1.0 indicate excessive unmodeled noise, that is, "they indicate that there is more variation between the observed and the model-predicted response patterns than would be expected if the data and the model were perfectly compatible." (Bond and Fox 2001 177)

Wright, Linacre, Gustafson and Martin-Lof (1994) provide a table of reasonable item mean square fit values and suggest infit and outfit values of 0.8 – 1.2 for high stakes tests, and 0.7 – 1.3 for 'run of the mill' tests. Values of the mean square statistics above 1.2 or 1.3 are considered as misfitting the model, whereas below 0.8 or 0.7 as overfitting the model. Overfit means close to a deterministic response string and too predictable by the Rasch model, but it is not considered a threat to the measurement

process. As explained by Wright et al. (1994) and Bond and Fox (2001), values of 1.3 (or 1.2) indicate 30% (or 20%) more variability than predicted by the Rasch model. Athanasou and Lamprianou (2002), Bond and Fox (2001) and Karabatsos (2000) suggest the same values as Wright et al. (1994).

Other researchers, such as Curtis (2004) and Glas and Meijer (2003) suggest using simulated data based on the estimated item parameters and then determining the critical values empirically. In such simulation studies researchers arbitrarily fix the Type I error rate (say 5%) and based on that they determine the cut-off value for the mean square statistics.

Whether simulation studies with a fixed Type I error are used, or the suggested cut-off values (which are rules of thumb) the decision as to which ones to use is arbitrary. Whichever method is used, misfit "should not be considered a 'have'/ 'not have' property but is always a matter of degree. As a matter of degree, the same misfit can be considered as too large or satisfactory depending on the aims of the measurement exercise" (Lamprianou 2006, 198).

The researchers believe that the amount of unmodeled noise present in a response pattern should be the criterion for identifying the degree of its aberrance and not the cut-off value for a fixed Type I error. In such a method researchers are willing to accept very different amounts of unmodeled noise as acceptable. For example, Petridou and Williams, (2007) used 1.72 for the outfit and Lamprianou (2006) 2.0 for both infit and outfit as cut-off scores. For the purposes of this study, given that classroom (low stakes) tests were used and following the suggestions of Wright et al. (1994), the conventional cut-off score of 1.3 for both infit and outfit statistics was used as the cut-off value which signaled that items needed to be reexamined.

With regard to identifying misfitting students, the researchers used six different cut-off scores for the two mean square statistics, varying from the conventional value of 1.3 to as high as 2.0, used by Lamprianou (2006). The values used were 1.3, 1.4, 1.5, 1.6, 1.8 and 2.0. For each cut-off score misfitting students in both tests were identified and then Chi square tests were performed to investigate possible associations between misfit in Test 1 and misfit in Test 2. All Rasch analyses were carried out with the help of WINSTEPS (Linacre 2005).

Results

Test 1 calibrations

The first calibration, in which the full set of the test data was used (27 items and 635 students), revealed two misfitting items, 10 and 2a (outfit > 1.5) and three slightly misfitting items, 7, 2c and 9a, (1.3 < outfit < 1.5). Also two of those items had infit of 1.50 (item 10) and 1.36 (item 7). The mean values of infit and outfit were 1.01 and 1.02 respectively.

Students' statistics revealed that 19 of them (3.0% of the sample) had outfit or infit (or both) higher than 2.7. Those students were considered badly misfitting and distorting the calibration process. They were therefore removed, leading to a second calibration with again the 27 items, but only with 616 students. The item estimates from this second calibration were anchored and used in the third and final calibration with the whole sample, including the 19 badly misfitting students, yielding ability estimates and fit indices for the full sample of students.

In the second calibration only three items were slightly misfitting, 10, 7 and 9a, with outfit statistics in the range 1.3 – 1.52 and infit statistics being 1.43 and 1.37 for the first two items. Item 9a (outfit = 1.32) was only marginally misfitting. Further investigation was conducted into the slight misfit of the 3 items.

Item 10 was the only item testing knowledge on straight line graphs, and it was decided not be removed because it included some very important skills in algebra, those of substituting values into a formula, plotting points and being familiar with the coordinate axes. It was an item of above average difficulty (0.48) and 52% of the students scored full marks while 20% of them scored no marks. Given that it was the only item testing this specific knowledge the slight misfit could have occurred because of special preference or special knowledge on straight line graphs.

Item 7 on the other hand was an item of approximately average difficulty (-0.14) on simple geometry. More than half, (57%) of the students, scored full marks whereas only 8% scored zero marks. This item required knowledge of basic angle properties. Misfit occurred because of some careless mistakes. It was considered too important to be removed.

Item 9a was a rather difficult item (difficulty 1.39) requiring knowledge of that if one or more terms in an equation are algebraic fractions, then the roots of their denominators are values of x that render the fractions undefined and must be excluded from the possible solution range. Only 38% managed to score full marks (2 marks) and 9% gave a half-correct answer scoring one mark. The remaining 53% of the students scored no marks. Misfit in this item occurred because of a few unexpected correct answers by students who most probably copied from more knowledgeable classmates.

These three items were not removed from the test for another reason. They were only slightly misfitting. A summary of the results of the Rasch analysis from the second calibration is given in Table 1.

< Insert Table 1 about here >

The range of student abilities was from -2.63 to 3.64, with a mean of 1.03 (SD = 1.17). Person reliability was 0.87. This index is an indication of the precision of the instrument

and shows how well the instrument can distinguish individuals. It is equivalent to Cronbach's alpha (alpha = 0.90). The student separation index was 2.61. This indicates the spread of person measures in standard error units, in this case in about 2.6 standard errors. A student separation index of 2.61 also indicates approximately four statistically distinct strata (strata = 3.81) of student abilities identified by the instrument.

The item estimates ranged from – 2.18 to 1.75 and the reliability index was 0.99. This index shows how well the items that form the scale are discriminated by the sample of respondents, in this case extremely well. The separation index is 11.40, indicating that the spread of item estimates is about 11 standard errors. The statistics of these items from the second calibration were then used for the third and final calibration which included the 27 anchored items and all 635 students.

Figure 1 shows the item-student map. One can see that most of the items are well targeted for students with abilities around and below the mean ability. Only 4 out of the 27 items are targeted at students with ability above the mean and those go up to about half a standard deviation above the mean. Overall, the bulk of the items (19 items) are well targeted for students with abilities ranging from 1 standard deviation below to half a standard deviation above the mean ability. Also, 6 items are targeted at students with ability of more than 2 standard deviations below the mean ability.

< Insert Figure 1 about here >

Given that the purpose of the test was to identify the weaker students the targeting of the items was considered very satisfactory.

Validity of Test 1

The qualitative approach to developing Test 1 and helping to ensure that it was valid for its purpose was described above and here the psychometric properties are set out. Table 2 shows the results of PCA on the standardised residuals.

< Insert Table 2 about here >

The variance explained by the measures (the dimension measured by the test) is 56.3% of the total variance. It is also about 17.5 times (34.7:2) the variance explained by the first factor extracted by PCA on the standardised residuals. The unexplained variance is 43.7% of the total variance in the data. The variance explained by this first factor is 7.4% of the unexplained variance (2 out of 27) and just 3.2% of the total variance in the data. This supports the hypothesis that there is no second dimension present in the data, therefore the test is unidimensional.

Furthermore, the scores on the test were compared with the final mathematics exam results of the students in the 3 schools. The correlation coefficients ($p < 0.01$ in all three cases) were: School 1: $r = 0.795$ ($N = 287$), School 2: $r = 0.704$ ($N = 37$), School 3: $r = 0.701$ ($N = 281$). The total number of students used for the correlation investigation was 605 (instead of the original 635) because 30 students were either asked to take the exams in September, or to repeat the year, as a result of too many unauthorised absences.

Comparisons of item estimates from two calibrations (Male-female subgroups)

Figure 2 shows the invariance plot for the two sets of item estimates subsets.

< Insert Figure 2 about here >

The points are closely scattered around the identity line, with only 3 out of the 27 items (approximately 11% of the items) clearly outside the 95% confidence interval (CI). In a binomial situation with 27 items, the probability of having three or more items outside the 95% C.I. is 0.15 ($p > 0.05$). Therefore three points outside the 95% C.I. is not a highly unlikely event if one has 27 items. The correlation coefficient between the two sets of item estimates was 0.975 which is extremely high.

These two results support the property of invariance of the Rasch model which indicates that the construct measured by the instrument has the same meaning to the two distinct groups.

All the evidence collected through the validation studies, together with the good fit of the test data to the Rasch model, support the hypothesis of a high degree of validity.

Test 2 calibrations

The calibration revealed one misfitting item (item 13: outfit = 1.78). The mean value of the infit and the outfit are 0.99 and 1.08 respectively.

Item 13 was the first non-multiple choice item and was an easy item (difficulty – 1.07) asking students to find the values of the constant p for which the quadratic equation $x^2 – 2x – 4p = 0$ has real roots. Six of the high scorers (with scores of 17 – 27) made careless mistakes on the item. Three of them lost 1 mark and three lost 2 marks (out of a maximum possible score of 4 marks). Even though the six students did not loose all the marks, their ability estimates were well above the item difficulty rendering their response to this item highly unexpected.

Once these six students were removed and the data reanalysed the outfit mean square value of item 13 dropped to 1.24, below the cut-off value. Therefore, item 13 was retained in the test because it was considered important for the unit's material and the high outfit value was caused by minor careless mistakes by only six students.

A summary of the results of the Rasch analysis is given in Table 3.

< Insert Table 3 about here >

The range of student abilities was from -3.30 to 3.21, with a mean of 0.25 (SD = 1.29). The reliability of student estimates was 0.82, not as high as the equivalent measures in the first test. However, given the fact that 12 out of the 16 items were multiple choice

items and the low stakes status of the test (a classroom test) the degree of reliability can be considered satisfactory. The student separation index was 2.13 and indicates approximately 3.2 statistically distinct strata of student abilities identified by the instrument.

The item estimates ranged from - 2.09 to 1.68 and the reliability index was 0.99. The separation index is 10.32, indicating that the spread of item estimates is about 10 standard errors. Figure 3 shows the item-student map.

< Insert Figure 3 about here >

The distributions of item difficulties and students' abilities are almost symmetrical indicating a very well designed test. The items are targeted at students with abilities from 2 standard deviations below to about one and a half standard deviations above the mean student ability. Also 8 items have difficulties above the students' mean ability and 8 items below.

Validity of Test 2.

As with Test 1 the involvement of professionals in the test development was outlined above. Here the quantitative data are considered.

Content validity questionnaire

Table 4 shows the responses of the eight experts to the content validity questionnaire.

< Insert Table 4 about here >

It is clear that all the experts agree or absolutely agree on almost all the statements. One of the experts disagreed with the format of the items, arguing that multiple choice items are not suitable for mathematics tests at this level. Also, two experts expressed their concern as to the time limits, arguing that the questions were perhaps too many to be

answered within 45 minutes. However, the administration of the test proved that there was no problem with the time given to the students to complete the test.

Table 5 shows the results of PCA on the standardised residuals.

< Insert Table 5 about here >

The variance explained by the measures is 65.8% of the total variance. It is also about 21 times the variance explained by the first factor extracted (30.9:1.5). The unexplained variance is 34.1% of the total variance in the data. The variance explained by this first factor is only about 9.4% (1.5 out of 16) of the unexplained variance and just 3.3% of the total variance in the data. Given these results one can conclude that there is no second dimension present in the data, therefore the test is unidimensional.

The scores on the test were compared with the final mathematics exam results of the students in the 3 schools. The correlation coefficients ($p < 0.01$ in all three cases) were:

School 1: $r = 0.840$ ($N = 259$), School 2: $r = 0.634$ ($N = 36$), School 3: $r = 0.751$ ($N = 141$)

The total number of students adds up to 436 (instead of the original 445) for the same reasons as in Test 1.

Comparisons of item estimates from two calibrations (male-female subgroups)

The data was again split into two groups based on gender. The two groups had sizes 186 (males) and 259 (females). Figure 4 shows the invariance plot for these two subsets.

< Insert Figure 4 about here >

The points are closely scattered around the identity line, with no items outside the CI, and that is a strong indication that invariance holds. Also, the correlation coefficient is 0.979, which is extremely high. These results support the property of invariance of the Rasch model.

Comparing ability estimates from calibrations of two different tests

The students' ability estimates from Test 2 were compared with the ability estimates from Test 1. The correlation coefficient between them was 0.706 ($p < 0.01$). When corrected for attenuation due to measurement error the correlation coefficient becomes 0.836. This high correlation strengthens even further the hypothesis that the two tests indeed measure the same ability, which was shown to be mathematical ability. It also strengthens our confidence in using the Rasch model, since the two tests, although both measuring mathematical ability, were targeted at different ability-level students. The first test was very easy and targeted at the lower ability students. The second test was targeted at about the mean student ability.

All of the above evidence, together with the fact that there was a good fit of the test data to the Rasch model, support the hypothesis of a high degree of validity for both tests.

Misfitting students

Table 6 shows the percentages of students identified as misfitting in the two tests by the outfit, the infit and the total (by outfit or infit or by both) for the various cut-off values.

< Insert Table 6 about here >

The percentages of misfitting students in the two tests are approximately the same except perhaps for one noteworthy fact. The outfit percentages are always larger for the second test, especially as the cut-off values increase. There are two reasons for this. Test 2 contained more multiple choice items where it is perhaps easier to make an unexpected mistake. At the same time Test 2 contained fewer items (16 instead of 27 in Test 1) and since the outfit is a mean square statistic, the impact of one unexpected response is larger.

Investigating whether aberrant response behaviour is a stable characteristic of students

Chi-square tests were then performed for each cut-off score investigating the null hypothesis of no association between misfit in Test 1 and misfit in Test 2. Table 7 shows for each cut-off score the observed frequencies in the four cells (with the row percentages given in brackets). For example, when 1.3 was used as the cut-off score, 92 (31.9%) of the fitting students in Test 1 were misfitting in Test 2 and 45 (28.7%) of the misfitting students in Test 1 were misfitting in Test 2.

The last two columns (the chi-square and the p-values) contain two numbers. The first is the one that corresponds to the test performed without using the continuity correction and the second in brackets to the test performed using the continuity correction.

< Insert Table 7 about here >

In two of the cases (for cut-off values of 1.4 and 1.6) the percentages of misfitting students in the second test from fitting and misfitting groups in the first test are identical. In all the other cases they are very close. Therefore, the results of the six chi-square tests are not even close to being significant and the null hypotheses of no association between the two variables at hand are clearly accepted.

Discussion

Various psychological and demographic characteristics of individuals have been reported to have an association with aberrant response behaviour. If indeed they have (or at least some of them do) then one would expect, as suggested by Smith (1986) and Lamprianou (2005), that an individual with an aberrant response pattern may exhibit such behaviour in other testing situations too. This means that aberrance may be a stable characteristic of individuals.

In the classroom setting maths test are more relevant, low-stakes, administered by the students' own comfortable-to-be-with teacher and one would perhaps expect less aberrance. This is a completely different context from the high-stakes tests administered in a much stricter and possibly a much stressful environment. At the same time, one would expect some type of aberrance to occur due to carelessness, sleepy behaviour, copying, cheating, plodding or guessing.

The findings of this study do not support Smith's and Lamprianou's suggestion that aberrance may be a stable characteristic of individuals. For the purposes of the study two classroom maths tests were used with a sample of 15-16 year old high school students in three different schools in Cyprus. The first test was administered to 635 students and the second to 445 of them. The Rasch PCM was used for the analyses of the data collected. Various studies verified that both tests had a high degree of reliability and validity.

Misfitting students in both tests were identified with the use of the infit and outfit mean square statistics for six different cut-off values. The hypothesis of no association between misfit in the one test and misfit in the other was investigated with the Chi square test and it was very clearly accepted with p-values much closer to 1.00 than to 0.05. It is concluded that misfit in the one test is not associated with misfit in the other among high school students taking classroom maths tests.

A couple of cautions should be made about this study. First, the test items used were mainly multistep mathematical problems with partial credit awarding for partial success instead of the usual dichotomous items found in the majority of studies on student aberrance. Perhaps it is easier to respond unexpectedly in dichotomous items, especially for high ability students, as reported by Petridou and Williams (2007). Where the answer is marked either right or wrong if a high ability student follows the correct

method (as expected) but gives the wrong answer (because of a careless mistake such as a miscalculation, or a miscopy of the right answer) he or she scores 0 and that signals his or her response as unexpected and probably the whole response string as aberrant (especially if the test is short). This is much less likely to happen with multistep problems. If such a mistake occurs, on the last stages of the solution process, the student will get most of the marks on that item and the answer will not be considered unexpected. Second the low stakes status of the tests linked to the administration procedure, with the familiar classroom setting may make the test takers feel more relaxed and perform more as expected than in a stricter and less familiar environment.

Concluding remarks

This study reports that misfit does not seem to be a stable characteristic of students taking classroom maths tests. This finding, together with Panayides' (2009) findings of no association between a large number of possible factors and misfit in the same setting, lead to the following intuitive conclusion: In classroom maths tests, although misfits do occur, they do not predict misfits in other tests and are not dependent on psychological or demographic characteristics of the test-takers. Therefore, high school maths teachers who test their students regularly should be aware that this kind of response behaviour does occur (perhaps leading to invalid estimates of their students' abilities) but should not be too concerned about it since they have many test results for their students and thus many ability estimates. Therefore, classroom maths teachers should identify and ignore ability estimates based on aberrant response patterns and use the ones from normal (expected) response patterns for the overall (or term) estimate of their students' abilities.

Finally, further investigation is suggested into other settings, with larger samples and with the use of multistep problems as well as dichotomous items. Given the

importance placed on the identification of misfitting students, the conclusion of this study, that aberrant response behaviour does not seem to be a stable characteristic of students, should be investigated further.

References

Athanasou, J. and I. Lamprianou. 2002. *A teacher's guide to assessment*. Sydney: Social Science Press.

Bond, T. G., and C. M. Fox. 2001. *Applying the Rasch model: Fundamental measurement in the social sciences*. New Jersey: Lawrence Erlbaum Associates.

Bracey, G. and L. M. Rudner. 1992. Person fit statistics: High potential and many unanswered questions. *Practical Assessment Research and Evaluation* 3, no. 7. http://pareonline.net/getvn.asp?v=3&n=7

Curtis, D. D. 2004. Person Misfit in Attitude Surveys: Influences, Impacts and Implications. *International Educational Journal* 5, no. 2: 125–144.

Douglas, G. A. 1990. Response patterns and their probabilities. *Rasch Measurement Transactions*, 3 no. 4: 75-77. http://www.rasch.org/rmt/rmt34a.htm

Emons, W. H. M., C. A. W. Glas, R. R. Meijer, and K. Sijtsma. 2003. Person-Fit in Order-Restricted Latent Class Models. *Applied Psychological Measurement* 27, no. 6: 459–478 .

Frary, R. B. and M. B. Giles. 1980. *Multiple-choice test bias due to answering strategy variation.* Paper presented at the annual meeting of the National Council on Measurement in Education, April 8–10, in Boston, Massachusetts.

Glas, C. A. W. and R. R. Meijer. 2003. A Bayesian approach to Person Fit analysis in Item Response Theory models. *Applied Psychological Measurement* 27, no. 3: 217-233.

Harnisch, D. L. and R. L. Linn. 1981. Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement* 18 no. 3: 133–146.

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement* 1, no. 2: 152-176.

Karabatsos, G. 2003. Comparing the Aberrant Response Detection performance of thirty six Person-Fit Statistics. *Applied Measurement in Education* 16, no. 4: 277–298.

Keeves, J. P. and G. N. Masters. 1999. *Advances in measurement in educational research and assessment*. Amsterdam: Pergamon.

Lamprianou, I. 2005. *Aberrant response patterns: Issues of internal consistency and concurrent validity.* Paper presented at the annual meeting of the American Educational Research Association, April 11–15, in Montreal, Canada.

Lamprianou, I. 2006. The stability of marker characteristics across tests of the same subject and across subjects. *Journal of Applied Measurement* 7, no. 2: 192–205.

Lamprianou, I. 2010. The practical application of Optimal Appropriate Measurement on empirical data using Rasch Models. *Journal of Applied Measurement* 11, no. 4: 409–423.

Lamprianou, I. and B. Boyle. 2004. Accuracy of measurement in the context of mathematics national curriculum tests in England for ethnic minority pupils and pupils who speak English as an additional language. *Journal of Educational Measurement* 41, no. 3: 239–259.

Levine, M. V. and F. Drasgow. 1988. Optimal Appropriateness Measurement. *Psychometrika* 53, no. 2: 161–176.

Linacre, J. M. 1998. Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement* 2, no. 3: 266–283.

Linacre, J. M. 2005. A user's guide to WINSTEPS [Computer program, version 3.57] Chicago: Winsteps.com.

Linacre, J. M. and B. D. Wright. 1994. Chi-square Fit Statistics. *Rasch Measurement Transactions* 8, no. 2: 360. http://www.rasch.org/rmt/rmt82a.htm .

Masters, G. N. 1982. A Rasch model for partial credit scoring. *Psychometrika* 47, no. 2: 149–174.

Mehrens, W. A. and I. J. Lehmann. 1991. *Measurement and Evaluation in Education and Psychology, (4th ed.)*. New York: Holt, Rinehart and Winston.

Meijer, R. R. 1996. Person-Fit Research: An Introduction. *Applied Measurement in Education* 9, no. 1: 3–8.

Molenaar, I. W. and H. Hoijtink. 1996. Person-Fit and the Rasch model, with an application to knowledge of logical quantors. *Applied Measurement in Education* 9, no. 1: 27–45.

Panayides, P. 2009. Exploring the reasons for aberrant response patterns in classroom maths tests. PhD diss. Durham University, UK.

Petridou, A. and J. Williams. 2007. Accounting for Aberrant test Response Patterns using multuilevel models. *Journal of Educational Measurement* 44, no. 3: 227–247.

Popham, W. J. 2000. *Modern Educational Measurement: Practical guidelines for educational leaders, (3rd ed.).* Boston: Allyn and Bacon.

Reise, S. P. and W. P. Flannery. 1996. Assessing Person-Fit on measures of typical performance. *Applied Measurement in Education* 9, no. 1: 9–26.

Rudner, L. M. 1983. Individual Assessment Accuracy. *Journal of Educational Measurement*. 20, no. 3: 207–219.

Smith, Jr., E. V. (2004). Evidence for the Reliability of Measures and Validity of Measure Interpretations: A Rasch measurement perspective. In *Introduction to Rasch Measurement*, ed. E. V Smith Jr, and R. M. Smith, **pages.** Minnesota: JAM Press.

Smith, R. M. 1986. Person Fit in the Rasch model. *Educational and Psychological Measurement* 46, 359–372.

Smith, R. M. 1990. Theory and practice of fit. *Rasch Measurement Transactions*. 3, no. 4: 78-79. http://www.rasch.org/rmt/rmt34b.htm.

Smith, R. M. 1996. Polytomous Mean-Square fit statistics. *Rasch Measurement Transactions* 10, no. 3: 516-517. http://www.rasch.org/rmt/rmt103a.htm .

Wright, B. D. 1977. Solving measurement problems with the Rasch model. *Journal of Educational Measurement* 14, no. 2: 97–115.

Wright, B.D. 1997. Measurement for social science and education: A history of social science measurement. http://www.rasch.org/memo62.htm

Wright, B. D., J. M. Linacre, J. E. Gustafson. and P. Martin-Lof. 1994. Reasonable mean square fit values. *Rasch measurement transactions* 8, no. 3: 370. http://www.rasch.org/rmt/rmt83b.htm .

Wright, B. D., and G. N. Masters. 1982. *Rating Scale Analysis*. Chicago: MESA Press.

Tables

Table 1. Summary of the results of the Rasch analysis for the mathematics test

	N	Estimate of mean (SD)	Range	Reliab.	Separ. Index	Infit msq mean (SD)	Outfit msq mean (SD)
Students	616	1.03 (1.17)	-2.63 to 3.64	0.87	2.61	1.06 (0.40)	0.97 (0.47)
Items	27	0.0 (1.21)	-2.18 to 1.75	0.99	11.40	1.01 (0.15)	0.97 (0.25)

Table 2. STANDARDIZED RESIDUAL variance (in Eigenvalue units)

		Empirical			Modeled
Total variance in observations	=	61.7	100.0%	100.0%	
Variance explained by measures	=	34.7	56.3%		59.1%
Unexplained variance (total)	=	27.0	43.7%	100%	40.9%
Unexpl var explained by 1st factor	=	2.0	3.2%	7.4%	

Table 3. Summary of the results of the Rasch analysis for Test 2

	N	Estimate of mean (SD)	Range	Reliab.	Separ. Index	Infit msq mean (SD)	Outfit msq mean (SD)
Students	445	0.25 (1.29)	-3.30 to 3.21	0.82	2.13	0.96 (0.67)	1.08 (0.79)
Items	16	0.0 (1.13)	-2.09 to 1.68	0.99	10.32	0.99 (0.08)	1.08 (0.23)

Table 4. Results of the analysis of the content validity questionnaire

Statements	Completely disagree	Disagree	Agree	Absolutely agree
The format of the questions is appropriate for the students	0	1	3	4
All the questions are clear and unambiguous	0	0	2	6
Students who know the answers have enough time to finish the test	0	2	4	2
All the important abilities and skills of the unit are assessed by the test	0	0	0	8
No irrelevant topics are included in the test	0	0	3	5
The test content is representative of the unit content as described in the curriculum	0	0	0	8

Table 5. STANDARDIZED RESIDUAL variance (in Eigenvalue units)

		Empirical		Modeled
Total variance in observations	= 46.9	100.0%		100.0%
Variance explained by measures	= 30.9	65.9%		65.3%
Unexplained variance (total)	= 16.0	34.1%	100%	34.7%
Unexpl var explained by 1st factor	= 1.5	3.3%		9.5%

Table 6. Percentages of misfitting students for the various cut-off values.

		Test 1			Test 2		
		Outfit	Infit	Total	Outfit	Infit	Total
Cut-off values	1.3	22.8	23.3	35.0	23.6	21.6	30.8
	1.4	19.2	17.2	27.9	20.0	16.4	27.4
	1.5	15.6	13.9	23.9	17.3	14.2	23,6
	1.6	12.6	11.3	20.0	16.2	12.4	22.0
	1.8	8.5	6.0	12.8	13.3	8.8	17.3
	2.0	6.1	4.6	9.9	11.7	6.7	14.4

Table 7. Chi-square tests for association between misfit in Test 1 and misfit in Test 2

		Test 2			
Cut-off	Test 1	Fitting	Misfitting	Chi-square	p-value
1.3	Fitting	196 (68.1%)	92 (31.9%)	0.514 (0.371)	0.474 (0.542)
	Misfitting	112 (71.3%)	45 (28.7%)		
1.4	Fitting	233 (72.6%)	88 (27.4%)	0.000 (0.000)	0.999 (1.000)
	Misfitting	90 (72.6%)	34 (27.4%)		
1.5	Fitting	261 (76.8%)	79 (23.2%)	0.104 (0.036)	0.747 (0.849)
	Misfitting	79 (75.2%)	26 (24.8%)		
1.6	Fitting	276 (78.0%)	78 (22.0%)	0.000 (0.000)	0.991 (1.000)
	Misfitting	71 (78.0%)	20 (22.0%)		
1.8	Fitting	323 (82.8%)	67 (17.2%)	0.034 (0.000)	0.854 (1.000)
	Misfitting	45 (81.8%)	10 (18.2%)		
2.0	Fitting	345 (86.0%)	56 (14.0%)	0.573 (0.281)	0.449 (0.596)
	Misfitting	36 (81.8%)	8 (18.2%)		

Figure 1. Students map of items

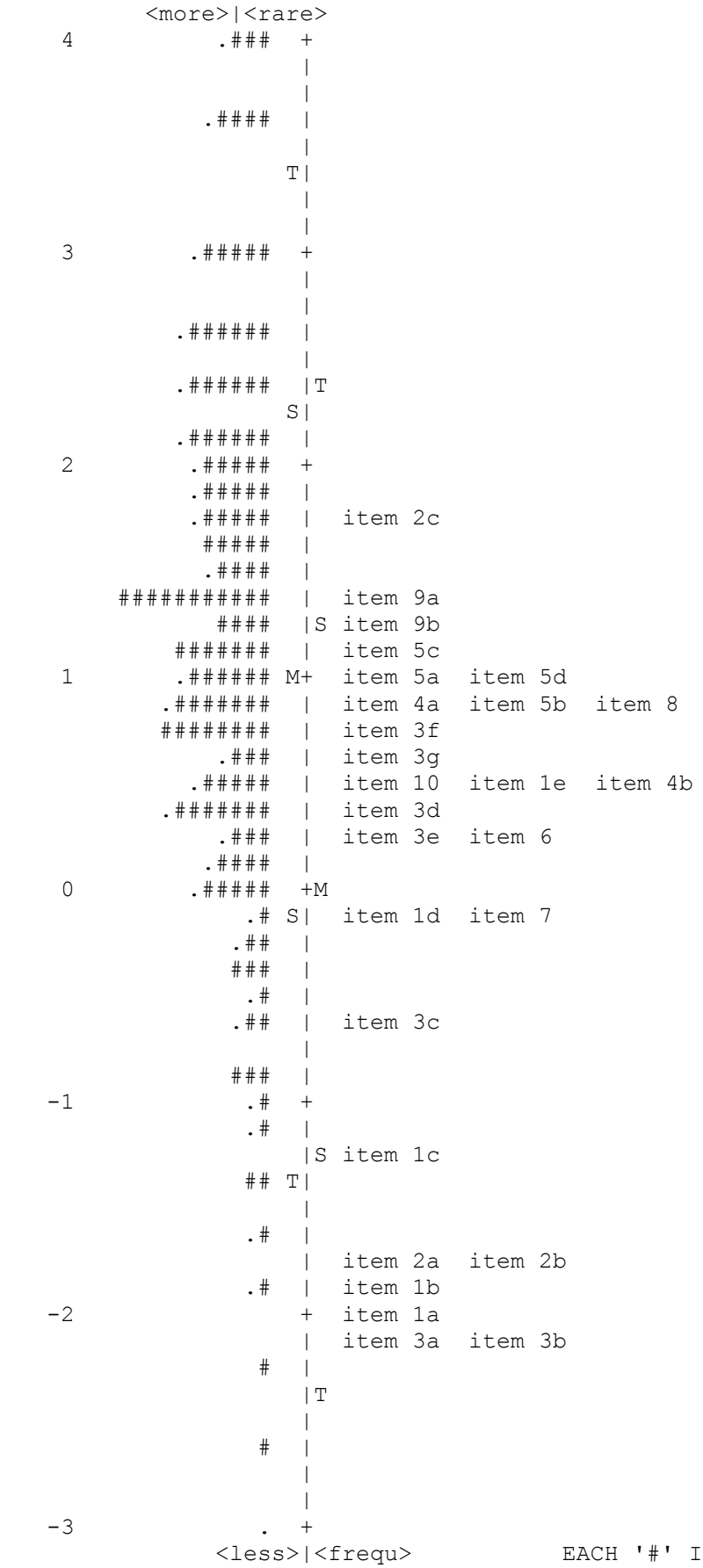

Figure 2. Invariance plot for Test 1(Item calibrations from male and female groups)

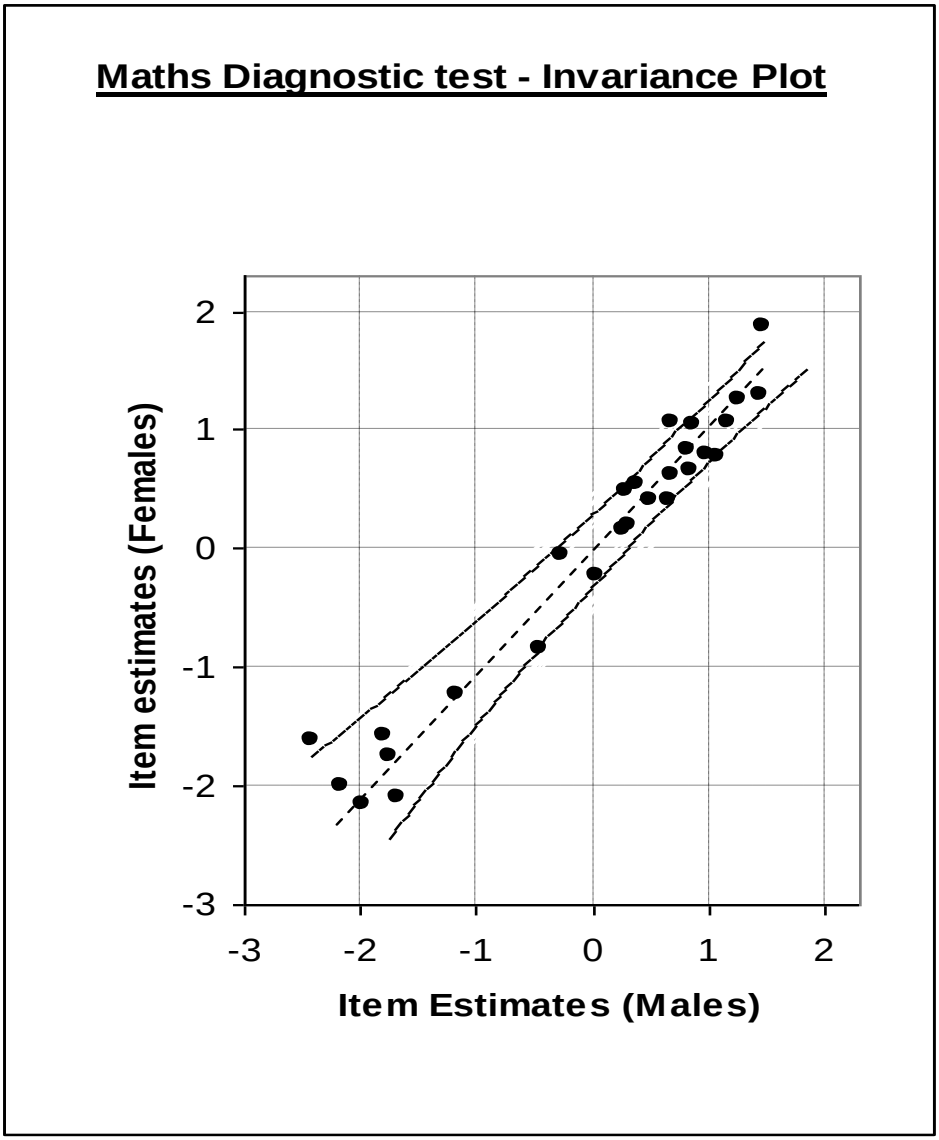

Figure 3. Students map of items

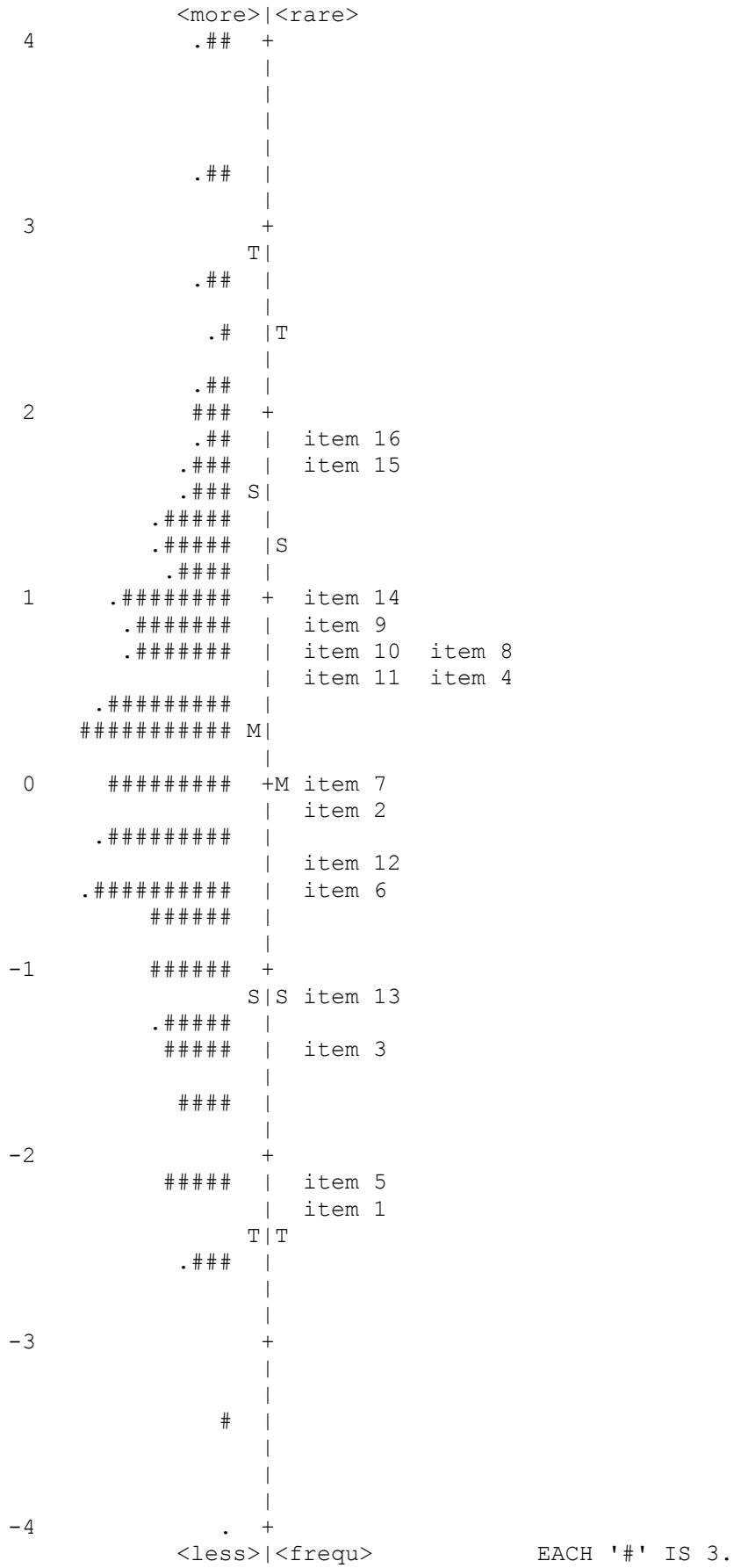

Figure 4. Invariance plot for Test 2 (Item calibrations from male and female groups)