

Durham E-Theses

Exploring the reasons for aberrant response patterns in classroom maths tests

Panayiotis Panayides

How to cite:

Panayides, Panayiotis (2009) Exploring the reasons for aberrant response patterns in classroom maths tests. Doctoral thesis, Durham University.

Use policy

The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a <https://etheses.durham.ac.uk/id/eprint/2042/> is made to the metadata record in Durham E-Theses
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Please consult the [full Durham E-Theses policy](#) for further details.

EXPLORING THE REASONS FOR ABERRANT RESPONSE PATTERNS IN CLASSROOM MATHS TESTS

by

Panayiotis Panayides

Submitted to the School of Education

For the Degree of

DOCTOR OF PHILOSOPHY

The copyright of this thesis rests with the author or the university to which it was submitted. No quotation from it, or information derived from it may be published without the prior written consent of the author or university, and any information derived from it should be acknowledged.

3 1 MAR 2009

DURHAM UNIVERSITY

2009



TABLE OF CONTENTS

Table of Contents	2
Acknowledgments	7
Declaration	8
Word count	9
Abstract	10
CHAPTER 1: INTRODUCTION	11
1.1 Educational Research	11
1.2 Measurement in the social sciences	25
1.3 This study	33
CHAPTER 2: REVIEW OF THE LITERATURE	37
2.1 Tests in general	37
2.1.1 Introduction	37
2.1.2 Classroom tests	38
2.1.3 Validity	42
2.1.4 Messick's modified view of validity	46
2.1.5 Factor analysis	48
2.1.6 Reliability	53
2.1.7 Item difficulty and discrimination in Classical Test Theory	61
2.1.8 Item Response Theory	64
2.2 The Rasch models	67
2.2.1 The Dichotomous model	67
2.2.2 Rasch model derived from objectivity	71
2.2.3 Assumption – model fit	76
2.2.4 Comparing the 2-P and the 3-P models with the Rasch model	80
2.2.5 Discrimination again: Is higher discrimination always better?	83

2.2.6	Rasch polytomous data	87
2.2.7	Criticisms of the Rasch model	91
2.2.7(i)	Rasch's different approach to the data-model relationship	91
2.2.7(ii)	The criticisms	92
2.2.8	Validity and Reliability addressed through the Rasch	101
2.2.8(i)	Validity	101
2.2.8(ii)	Reliability	111
2.3	Appropriateness measurement	114
2.3.1	Introduction	114
2.3.2	Possible factors associated with misfit	116
2.3.3	Person Fit statistics	124
2.3.4	Infit and outfit mean square statistics	129
2.3.4(i)	Introduction	129
2.3.4(ii)	Critical values for the infit and outfit mean Square statistics	132
2.3.4 (iii)	Uses and criticisms of the infit and outfit mean square statistics	135
2.3.5	Misfit as a threat to measurement	139
CHAPTER 3: METHODOLOGY		141
3.1	Ethics	142
3.2	Phase 1	143
3.2.1	The maths test	144
3.2.2	Selection of the Rasch models and fit statistics	145
3.2.3	Validity and reliability of the maths test in phase 1	149
3.2.4	Test Anxiety Inventory (TAI)	150
3.2.5	Assessment of Attention Deficit Hyperactivity Disorder	152
3.2.6	The investigation of factors associated with misfit	153
3.3	Phase 2	156

3.3.1	The first maths test (The Diagnostic Test) in phase 2	156
3.3.2	The second maths test	158
3.3.3	The maths self-esteem scale	160
3.3.4	Shorter version of the Test Anxiety Inventory	162
3.3.5	Predictive validity and internal consistency of scores of fitting and misfitting students	163
3.3.6	The interviews	164
3.3.7	Infit and outfit investigations	167
CHAPTER 4: RESULTS		168
4.1	Phase 1 results	169
4.1.1	The maths test	170
4.1.2	Reliability and validity of the test	179
4.1.3	Misfitting students	191
4.1.4	Test Anxiety Inventory (TAI)	193
4.1.5	Validity of TAI	193
4.1.6	TAI calibrations	199
4.1.7	Misfitting students	206
4.1.8	Assessment of Attention Deficit Hyperactivity Disorder (ADHD) characteristics	207
4.1.9	Other factors considered	212
4.1.10	Investigation of possible factors associated with misfit	222
4.1.11	Comparing the internal consistencies of raw scores of fitting and misfitting students	226
4.2	Phase 2 results	234
4.2.1	The first maths test and its calibrations	234
4.2.2	Reliability and validity of the test	243
4.2.3	Misfitting students	248
4.2.4	Math Self-esteem scale (MSES)	250
4.2.5	Reliability and validity of the MSES	261
4.2.6	The second maths test (test on quadratic equations)	268
4.2.7	Reliability and validity of the 2 nd test	277

4.2.8	Misfitting students	285
4.2.9	The shorter version of the TAI	287
4.2.10	Reliability and validity of the TAI	293
4.2.11	Misfitting students	299
4.2.12	Detecting multidimensionality: PCA of Rasch standardised residuals or PCA of raw scores?	301
4.2.13	Investigation of possible factors associated with misfit	308
4.2.14	Predictive validity of the test scores of fitting and misfitting students	315
4.2.15	Comparing the internal consistency of responses for fitting and misfitting students	319
4.3	The interviews	324
4.3.1	Reasons for misfit	325
4.3.2	Possible artifact (inflated outfit)	339
4.4	Investigating outfit and infit	343
4.4.1	The effect of test length on the outfit of a response string with one unexpected response	343
4.4.2	A formula for the contribution of an unexpected response to the outfit	353
4.4.3	Investigating the effect of unexpected responses to the outfit for items	357
4.4.4	The number of items with 'less likely' answers for which infit exceeds the cut-off value	360
CHAPTER 5: CONCLUSIONS		370
5.1	The procedure	370
5.2	Factors investigated for possible association with misfit	373
5.3	Results of the interviews	376
5.4	Is misfit an inherent characteristic of students?	377
5.5	Predictive validity	378
5.6	Internal consistency	378

5.7	Investigation of outfit and infit	379
5.8	Limitations of the study	382
5.9	Final comments	383
References		386
Appendices		403

Acknowledgements

I have very much appreciated the opportunity to conduct my PHD at the School of Education of the Durham University. Even though I have visited the university only a few times the atmosphere there was very friendly and supportive and I have always enjoyed the time I spent there.

I feel particularly lucky and honoured to have done my research under the supervision of Professor Peter Tymms. He is a highly gifted academic who has made me feel part of the team right from the start. His guidance and recommendations have always been highly constructive.

Even at times when I would send him numerous emails with pieces of my work for him to have a look at and queries, he would always find the time to answer promptly, despite his heavy schedule, and his comments and suggestion have been not only constructive throughout my research but also motivating and intellectually challenging.

I am very grateful for his very reliable and highly competent guidance.

Declaration

I declare that no part of the material in this thesis has previously been submitted by me for a degree in this or any other university

The copyright of this thesis rests with the author. No quotations from it should be published without his prior written consent and information derived from it should be acknowledged.

Word count

The word-length of this thesis is approximately 96,400 words, including tables and figures.

Contents, Abstract, Declaration, References and Appendices are excluded from this word count.

Abstract

This study has focused on the investigation of the reasons for aberrant response patterns in classroom maths tests.

Data were collected from high schools in Cyprus over two academic years. The assessment instruments used included: three Maths Tests, a Test Anxiety Inventory (TAI) and a shorter version of it, an Attention Deficit Hyperactivity Disorder (ADHD) scale and a Maths Self-Esteem Scale.

Results showed no associations between any of the factors investigated and misfit when tests with polytomous items were used. Factors investigated included: student and teacher gender, item order, different schools, different teachers, ability, test anxiety, ADHD, maths self-esteem, motivation, language competency, interest in maths, private tuition in maths, study time and class revision. This finding has led to the investigation whether misfit is an inherent characteristic of students and the conclusion that it is not.

The only factors that showed some association with misfit were ability ($p = 0.022$), the interaction of gender with test anxiety ($p = 0.018$) and different teachers ($p = 0.027$), and the first two were only for the test containing 12 (out of 16) dichotomous items. Further investigation of these factors is suggested.

Analyses of interviews of 21 misfitting students showed that the main reason given for unexpected responses among high ability students was, as expressed by them, carelessness and among low scorers prior knowledge and to a lesser degree cheating and special preference.

The two mean square statistics, infit and outfit were also investigated, and an explanation is given for why high infit is considered more of a threat to measurement than high outfit. The researcher finally argues that students with misfitting patterns with high outfit values should not be considered as invalidly measured without further investigation. Similarly, items with high outfit should not be considered as malfunctioning and removed without further investigation.

CHAPTER 1: INTRODUCTION

This chapter consists of 3 parts. In the first part educational research is defined, followed by a brief historical review and references to the debate about methods used and the criticisms of educational research. The second part discusses measurement in the social sciences with a special focus on Rasch measurement and appropriateness measurement. Finally, the last part provides a brief description of the purposes of this study.

1.1 Educational Research

The word research comes from the French word ‘recherché’ which means ‘to investigate thoroughly’.

Scientific research is systematic, controlled, empirical and critical investigation of natural phenomena guided by theory and hypotheses about the presumed relations among such phenomena.

(Kerlinger, 1986, p. 10)

Kerlinger (1986) emphasises two points from his definition of scientific research. First, ‘systematic’ and ‘controlled’ meaning that scientific investigation is so ordered that investigators can have critical confidence in their research outcomes. Second, scientific investigation is ‘empirical’ meaning that if scientists believe that something is so, they must somehow put their belief to a test outside of themselves. In other words “subjective belief must be checked against objective reality” (p. 11).

Social sciences (such as education, psychology, sociology, anthropology and philosophy) are a “branch of science that deals with the institutions and functioning of human society and with the interpersonal relationships of individuals as members of society” (Webster’s Ninth New Collegiate Dictionary, p.1119). On top of that social science is concerned with the whole person and his/her mental, spiritual, physical and emotional development.

Educational research is defined by the Higher Education Funding Council of England as “an original investigation undertaken in order to gain knowledge and understanding” (HEFCE, 1999, p. 261) whereas the British Educational Research Association, BERA, (2000) defines two main thrusts to educational research.

These are:

- To inform understandings of educational issues, drawing on and developing educational theory, and in some cases theory from related disciplines (e.g. sociology, psychology, philosophy, economics, history etc).
- To improve educational policy and practice, by informing pedagogic, curricular and other educational judgments and decisions.

Mortimore (2000) discusses the following major tasks of educational research:

- To conceptualise, observe and systematically record events and processes to do with learning.
- To analyse such observations in order to describe accurately their conditions, contexts and implications.
- To publish accounts of all that is known about a particular topic under consideration, drawing on existing theory from one of the disciplines that contribute to our field, from educational theory itself, or from emerging theory that will itself be aided by the work.

The main purpose, in Mortimer’s view, is to further educational improvement. Educational research can do this most easily through the advancement of trustworthy knowledge about education.

According to McGaw (1997) educational research includes:

- Basic research (e.g. study of the motivation of young children).
- Applied research which sets such an inquiry in the context of a particular problem (e.g. how do teachers evoke greater motivation from 6 year old pupils?).
- Experimental development of the research ideas (e.g. offering pupils greater choice, or independent counselors and evaluating the impact on motivation).

- A radical approach to research which stems from the “blast of deconstruction which postmodernist questioning has landed on the kinds of truth claims pursued by the research traditions” (Brown, 1997, p. 81).

This last view is the one which, according to Mortimore (2000), even though it challenges the assumptions we make about ourselves and may be hard to grasp in relation to existing paradigms, should not be ignored. The researcher endorses Mortimore’s view. In fact a good example, and one directly related to this study, is Rasch’s pioneering work with his model, with which he challenged the traditional data-model relationship.

Mortimore (2000) points out that the scope of educational research seems enormous and ranges from studies of the learning of babies and young children to the life long learning of adults. It includes anything to do with the educative process and many topics within health, childcare and delinquency. It may focus on places where education takes place (schools, playgrounds, libraries or homes) or on people (pupils, teachers, childcare workers, parents, support staff, chief education officers or civil servants).

The various definitions of educational research quoted in this introduction do not contradict each other, they rather complement each other. It is the researcher’s opinion that a more condensed and formalized statement could be ‘educational research is a systematic investigation into educational issues aiming at the better understanding of these issues, the advancement of existing knowledge and the improvement of educational policies’.

History of educational research

De Lansdheere (1993) gives a historical review of educational research and in pages 4-5 lists the following late 1800s events which he associates with the birth of modern educational research:

- 1885 Ebbinghaus’s study on memory, which drew the attention of the education world to the importance of associations in the learning.
- 1888 Binet published his *Etudes de Psychologie Expérimentales* (Studies in experimental psychology)

- 1890 The term 'mental test' was coined by Cattell.
- 1891 Stanley Hall launched the review *Pedagogical Seminary*.
- 1894 Rice developed a spelling test to be administered to 16,000 pupils.
- 1895 The National Society for the Scientific study of Education was founded in the United States.
- 1896 In Belgium, Schyten published a report of his first educational research study on the influence of temperature on school children's attention.
- 1897 Thorndike studied under James at Harvard and there discovered the works of Galton and Binet. Ebbinghaus published his so-called completion test to measure the effect of fatigue on school performance.
- 1898 Lay suggested distinguishing experimental education from experimental psychology. Binet and Henri condemned traditional education in their book *La Fatigue Intellectuelle* and indicated the need for experimental education.
- 1899 Schyten opened a pedagogical laboratory in Belgium to study experimentally, among other things, group teaching methods.

De Lansdheere (1993) continues his historical review with the 20th century.

During 1900 – 1930 most educational research was quantitatively oriented and geared to the study of effectiveness. In an attempt to obtain sufficient validity of measurement for the complexity of most phenomena, researchers have achieved many statistical advances.

- In 1904, Spearman published his analysis of a correlation matrix to sustain his two-factor theory of intelligence and factor analysis began to emerge. The same year also marks the appearance of the first textbook in measurement theory 'An Introduction to the Theory of Mental and Social Measurement' by E. L. Thorndike.
- In 1908 Gosset, under the name of Student, showed how to measure the standard error of the mean and the principle of the t-test was formulated.
- Group testing began in England in Galton's laboratory in 1905 and Burt and Spearman assisted him.
- In 1911 the US National Education Association approved the use of tests for school admission and final examinations.

- The 1918 Yearbook of the National Society for the study of Education was entirely devoted to the measurement of educational products.
- In 1928 about 1,300 standard tests were available in the US and by the 1930s item formats, order of items, parallel forms, scoring stencils and machine scoring, norms, reliability and validity were fully developed.
- According to Dubois (1970) measurement theory began to blossom in the 1930s. In 1935 the journal 'Psychometrika' was founded, followed in 1941 by 'Educational and Psychological Measurement' and in 1947 by the British 'Journal of Statistical Psychology'.

The Second World War and the years immediately after brought educational research activities in European countries to a stand still. In the US, Australia and Sweden things were different.

Allen and Yen (1979) claim that although research into methods of psychological measurement continues most of the foundations for present day measurement theory were completed by the 1950s.

During the first half of the 1960s, in wealthy countries educational research received, for the first time, the support necessary for it to have a significant impact, especially in the US. At the same time large private foundations also began to sponsor educational research on a large scale.

Scientific achievement in the field of education in the 1960s, according to De Lansdheere (1993) include amongst others:

- New concepts of criterion-referenced testing
- Formative and summative evaluation
- Research on teacher effectiveness
- Adult education
- Research in methods of early education
- Social aspects of learning aptitudes
- Development in research methodology

It was towards the end of the 1960s and the beginning of the 1970s that people began to react against the dominant quantitative methods that have been traditionally used, and

from that reaction qualitative methods emerged (for example Campbell, 1974; Cronbach, 1974; Hargreaves, 1967).

Quantitative vs qualitative methods

Quantitative research is, as the term suggests, concerned with the collection and analysis of data in numeric form. It tends to emphasise relatively large-scale and representative sets of data, ... Qualitative research, on the other hand, is concerned with collecting and analysing information in as many forms, chiefly non-numeric, as possible. It tends to focus on exploring, in as much detail as possible, smaller number of instances or examples which are seen as being interesting or illuminating, and aims to achieve 'depth' rather than 'breadth'.

(Blaxter, Hughes and Tight, 2001, p. 64)

So, quantitative methods usually deal with statistical techniques on large scale data (sometimes small scale numerical work with ANOVA tests or other techniques can be dealt with in quantitative research) whereas qualitative methods deal with exploring in detail, with non-numerical analyses small numbers of cases.

According to Blaxter et al (2001), there have been ongoing debates in recent years regarding the relative merits of quantitative and qualitative methods with some social scientists supporting the one and others supporting the other. These debates are referred to as "paradigm wars" and the participants in these as "warriors" by Tashakkori and Teddlie (1998, p. 4). "Warriors" like Lincoln and Guba (1985) and Smith and Heshusius (1986) have claimed an incompatibility of the two different methods with the last suggesting giving up the dialogue between the two camps because further dialogue was unproductive. This point of view was called the incompatibility thesis (Tashakkori & Teddlie, 1998, p. 4).

Social scientists who attempted to make peace between the warriors of the two camps (for example Howe, 1988; Reichardt & Rallis, 1994), presented the compatibility thesis and adopted the view that whatever philosophical and/or methodological approach works for the particular research problem under study should be used.

Brewer and Hunter (1989) note that most areas of research in the social and behavioral sciences now use multiple methods and with the tremendous growth of social sciences since the fifties “there is now virtually no problem area that is studied exclusively within one method” (p. 22).

Tashakkori and Teddlie (1998) argue that mixed methods should be used because both quantitative and qualitative methods have been used for many years in empirical research, funding agencies have accepted them and both have led to generally accepted results thus influencing policies. What they are implying is that since both methods have been used successfully over the years and in many cases they seem to complement each other, there is no reason why they could not both be used in the same investigation.

The concept of mixing different methods probably originated in 1959, when Campbell and Fiske used their Multimethod Multitrait matrix to examine multiple approaches to data collection in a study. This encouraged others to mix methods and soon qualitative methods, like interviews, were combined with traditional surveys. Recognizing that all methods have limitations, researchers felt that biases inherent in one method could be neutralized by other methods.

Creswell (2003) states:

For example, the results from one method can help develop or inform the other method... Alternatively, one method can be nested within another method to provide insight into different levels or units of analysis.... Or the methods can serve a larger, transformative purpose to change and advocate for marginalized groups, such as women, ethnic/racial minorities, members of gay and lesbian communities, people with disabilities, and those who are poor... These reasons for mixing methods have led writers from around the world to develop procedures for mixed method strategies of inquiry. (pp 15-16).

Westmarland (2001), who in describing research methods adopted for feminine use, supports the use of mixed methods and emphasises (like Creswell, 2003) the complementary role of each method to the other by noting that although a survey (the quantitative approach) may be the best way to discover the prevalence of problems, interviews (the qualitative approach) will help to understand better women's experiences and theorise these experiences with a view towards social change.

‘For example, a survey can tell us that women working outside the home generally get paid less than men, but does not explain how this makes women feel and how it affects their lives’ (Westmarland, 2001, par. 27)

The importance of educational research and how its value can be enhanced

Stanley (1991) argues that research in education is vital if the education community is to rise to the challenges brought by increased participation and equity in the context of microeconomic reform and award restructuring.

Educational research is intellectually demanding and at times very frustrating. In the absence of good research, opinion and superstition prevail. Even in the presence of good educational research the same conditions can apply.

While some outcomes of educational research are not what people wish to hear, there is greater likelihood of change to the extent that sound data are available. For example, it is much harder for someone to assert that educational standards are falling, if there are good comparable data that refute this.

Mortimore (2000) places emphasis on the importance of educational research by listing some of its successes. These, amongst others include:

- Radical approaches of the early researchers in special education who showed the way to use knowledge to improve the lives of people who had been written off by society.
- Studies devoted to uncovering lack of equality in the UK educational system. Studies of social class, gender and race issues which have changed the way pupils are treated.
- An Inner London Education Authority study of women’s career in teaching showed that the proportionate success of women competing for promotion was higher than their male counterparts but because in terms of absolute numbers women applicants were fewer, men appeared to be more successful. Revealing the reality of these data encouraged more women to apply for promotion and succeed.

Given the importance of Educational research Mortimore (2000) also suggests ways in which its value can be enhanced:

- Researchers need to work within the professional and ethical BERA codes and revise such codes regularly.
- Everything published should meet the criteria set by research education authorities.
- Conflicting research results and methodological antagonism should be acknowledged and accommodated.
- Invest in learning. New techniques are being developed and should be included in the researchers' repertoire.
- Researchers should develop their information handling skills to a much more sophisticated level, given the volume of material that is available.

These ways suggested by Mortimore can be used as guidelines for enhancing educational research with emphasis on the training of new researchers on following them.

The BERA values were presented neatly and in a very condensed form in the presidential address of Jean Rudduck in 1995 (as cited in Mortimore, 2000, p.20) as "respect for evidence, respect for persons, respect for democratic values and respect for the integrity of our acts at every level of research enterprise".

Criticisms of educational research

One of the major criticisms of educational research is that researchers present their findings "in a form or medium which is largely inaccessible to a non-academic audience and lack interpretation for a policy-making or practitioner audience" (Hillage, Pearson & Tamkin, 1998).

Three more criticisms are described in detail by Mortimore (2000). These are:

- Educational research is frequently biased. However, bias is an ever-present danger for all researchers to be aware of and to guard against.
- It is perceived as threatening, especially by politicians and social workers. They seem to resent the authority that comes from a systematic investigation; the more so if research findings contradict received wisdom

or challenge policy. Other researchers can also feel threatened by work which contradicts their findings.

- The relative poor standing of education in relation to other subjects and of educationists in relation to their peers in the sciences, law or even other social sciences.

Another criticism, mentioned by Shavelson (1988), on top of the perception that educational research is threatening, is the questioning of policymakers and practitioners on the contribution of social science research to policy and practice. Shavelson however, argues that the perception that educational research does not significantly contribute to practice is inaccurate.

This perception grows out of policymakers and practitioners who get disappointed when their own unrealistic expectations that educational research should directly and immediately influence policy or practice the same way physical or medical science research do, are not met.

These expectations, according to Shavelson (1988), rest on the following unrealistic conditions:

- Research would have to be relevant to a particular issue and be available before a decision has to be made.
- It should provide clear, simple and unambiguous results.
- It would be known and understood by policymakers and practitioners and not cross entrenched interests.
- Recommendations from research would be implemented within existing resources.
- Research findings would lead to choices different from those that decision-makers would have otherwise made.

On a similar note, Campbell (1969) argues that reform administrators believe that specific social reforms advocated are certain to be successful. "Trapped Administrators have so committed themselves in advance to the efficacy of the reform that they cannot afford honest evaluation. For them favorably biased analyses are recommended ..." (p. 426).

What modern nations need, according to Campbell (1969), is readiness for an experimental approach to social reform in which new programs are designed to cure

specific social problems, tried out and if they are found to be ineffective they are modified or discarded. He then suggests a change in the political postures which will further a truly experimental approach to social reform.

One simple shift in political posture which would reduce the problem is the shift from the advocacy of a specific reform to the advocacy of the seriousness of the problem, and hence to the advocacy of persistence in alternative reform efforts should the first one fail.

(Campbell, 1969, p. 410)

The Western Australian Institute for Educational Research, WAIER, (1991) adds to the criticisms that most educators today point to good educational research being undertaken at the various tertiary institutions, and some other research centres, but little evidence is found of research effort impacting on changing the nature of what is happening at the classroom level.

WAIER (1991) suggests that communication between university researchers and classroom teachers should be improved thus disseminating the research findings to ensure translation into more effective practices at the classroom level. Educational researchers and classroom teachers should work together on matters of educational significance and should combine the research expertise of university academics with the practical knowledge of classroom teachers.

Where classroom participants join educational researchers as the doers of research, a greater degree of change and improvement at the classroom level is likely to follow.

(WAIER, 1991, p. 44)

Another criticism of educational research can be found in two 1998 publications. One was James Tooley's study entitled 'Educational Research, a critique' and the second a report by the Institute of Employment Studies (IES). Both publications claimed that the £65 million spent by Government on funding educational research was wasted, since much of the research was of dubious quality.

Tasker and Packham (1998) discuss the findings of the two reports adding that those findings were considered by the Minister of Higher Education who suggested a shift in Government policy: educational research should be concentrated in 10 to 20 centres of research excellence directing their work towards what works best in the classroom.

Tasker and Packham (1998) comment that the significance of the issue does not lie in the predictable attack on academics in university education departments but in the highhandedness of the educational policymakers. They also question the objectivity of the two reports since they were both commissioned by the Government (Tooley's by OFSTED and IES's by the Department for Education and Employment); they were limited in scope and carried out over a period of only a few months.

They conclude their article by emphasizing the dangers of extending the narrowing down tendency of the Government's educational policy to educational research.

If confined to a few centres of research (selected by the Government) and directed in the selection of subject matter, educational research will fall into place in a centrally controlled national education system subject to greater Government control. Academic freedom will gradually wear out as researchers working in universities not considered as centres of excellence will be silenced and those who do obtain funding will not have the opportunity to research into what they think are worthwhile.

The current status of allocation of funds in the UK

The Research Assessment Exercise (RAE) is a periodic UK exercise undertaken approximately every 5 years on behalf of the 4 UK higher education funding councils. These are the Higher Education Funding Council of England (HEFCE), the Scottish Higher Education Funding Council (SHEFC), the Higher Education Funding Council of Wales (HEFCW) and the Department for Employment and Learning for Northern Ireland (DELNI).

RAE aims to assess the quality of research activity in a range of subject areas (called Units of Assessment, which often represent the different university departments). A subject specialist peer review panel ranks each unit of assessment and these ranks are used to inform the allocation of quality weighted research funding each higher institution receives from the national funding council.

According to Wikipedia (accessed 03/07/2008) the RAE has been criticised by the University and College Union in that it has lead to the closure of departments with strong research profiles and healthy student recruitment. They also blame RAE for job

losses, discriminatory practices, the narrowing of research opportunities and the undermining of the relationship between teaching and research.

Roberts (2003) in his Review of Research Assessment which was commissioned by the UK funding bodies and known as the Roberts Report recommended changes to research assessment partly in response to the criticisms. This report was taken under consideration by the House of Commons Science and Technology select committee who concluded that RAE had positive effects and that a marked improvement in university excellence was evident. Finally they proposed a reformed RAE based on Roberts' recommendations.

It was announced in the 2006 Budget (according to Wikipedia) that after the 2008 exercise a system of metrics would be developed in order to inform future allocations of research funding.

Educational research and policymakers

Educational researchers cannot ignore the democratically elected government of the country which has the power to control many aspects of researchers' lives, and is many times critical to their work. Mortimore (2000) suggests that despite the criticisms researchers must continue to seek ways to work with the government by:

- Maintaining channels of communication through which they can dispute what they believe to be wrong judgments.
- Collaborating on appropriate projects, such as the establishment of a National Research Forum.
- Listening to, and taking seriously, the government's legitimate criticisms of their work.

At the same time researchers must:

- Generate their own research topics.
- Evaluate government actions and policies.
- Use academic freedom to question and dispute, responsibly and positively, any matter on which they have expertise or knowledge gleaned from their research.

Shavelson (1988) argues that the reason for educational research being sometimes ineffective is a “mismatch of mindframes” (p. 9) because the researchers’ mindframe does not easily translate into the policymakers’.

Research bureaucrats are people who work in agencies, usually Government agencies, and are responsible for commissioning and overseeing research and translating it into information useful for policymakers.

If research is to have an impact on policy, research bureaucrats are the people who would most probably be aware of research and find it useful in their job.

Shavelson (1988) lists 5 criteria which bureaucrats use in judging the usefulness of a study, and educational researchers should be aware of. These are:

- a) Technical quality. (the most important criterion)
- b) Recommended actions that policymakers can do something about.
- c) The fit with the bureaucrats’ prior knowledge.
- d) Whether a study challenges accepted truth.
- e) Whether a study is relevant to an issue.

Therefore, as Yates (2002) suggests, a researcher needs to think about who will be judging the successfulness of the research, what their criteria are and what they will do when they judge it.

And, as Mortimore (2000) concludes, educational researchers should do in the future what they have been trained to do; ask difficult questions, generate, through their research, new knowledge, formulate new theories and speak up for what they believe is right.

One of the concepts of educational research which always concerns educators, researchers and policymakers in the social sciences is that of measurement. The role of measurement is to provide decision makers with accurate and relevant information. Educators, and more generally behavioural scientists, have been treating measurement as a necessary component in both research and practical decision making.

1.2 *Measurement in the social sciences*

Measurement implies a much broader concept than a test.

We can measure characteristics in ways other than giving tests. Using observations, rating scales, or any other device that allows us to obtain information in a quantitative form is measurement.

(Mehrens and Lehmann, 1991, p. 4)

Stevens (1946) defined measurement as the assignment of numbers to objects according to a rule; therefore some sort of measurement exists at the nominal, ordinal, interval and ratio levels.

In spite of Stevens's personal claim to the contrary, we know that ratio-level measurement is likely to be beyond our capacity in the human sciences, but most of us do well enough by regarding the data that we have collected as belonging to interval-level scales.

(Bond and Fox, 2007, p.2)

According to Bond and Fox (2007), over the last century educators, psychologists and generally researchers in the social sciences have focused on the application of sophisticated statistical procedures to their raw data. In fact they were too narrowly focused on statistical analyses "and not concerned nearly enough about the quality of the measures on which they use these statistics" (p.2).

Many of the data collected in the social sciences, like Likert scales or test scores, are mistakenly regarded as belonging to the interval-level scales.

Bond and Fox also claim that this persistent reliance on raw scores originated from Stevens' definition of measurement.

Adhering to Stevens' definition limits our thinking to the level of the raw data. According to Bond and Fox, under the mistaken belief that they are measuring psychologists, educators and researchers in the social sciences describe the raw data at hand. They report how many people answered an item correctly or how many items were answered correctly by people, thus assigning scores. Bond and Fox (2007) argue

that these do not constitute measurement but are mere descriptions. Wright (1994) argues that when giving out a maths test we are not interested in how many (the raw score) or which items a person answered correctly but how much maths the person knows.

Michell (2003) also refers to Stevens' definition of measurement by stating that "it misses the mark entirely" (p. 304). He very convincingly argues that while measurement involves objects and events, it focuses on attributes of objects and events and to be more precise on quantitative attributes of things. The standard definition of measurement neglects the concept of a continuous quantitative attribute and by doing so, "misses the concept at the heart of the matter" (Michell, 2003, p.305).

The use of raw scores as measures raises a couple of important issues:

- Wright (1999) argues that raw scores are bound to begin at 'zero score' and end at maximum score. Does zero score mean no ability at all and maximum score the maximum possible ability? Surely the continuous quantitative attribute of people we are trying to measure cannot have any boundaries. There should be no maximum ability or complete lack of ability.
- If a person A obtains a score on a test (say 20) and person B obtains double that score (say 40) we could not say that person B has twice as much of the trait being measured. "To make such a statement would require that one assumes a score of zero to actually represent no amount of the characteristic. In general, if a person received a score of zero on a spelling test, we could not interpret that score to mean that the person had no spelling ability. The same is true for any other test" (Mehrens and Lehmann, 1991, p.211).
- More importantly however, does the difference between scores of 50 and 51 represent the same difference in abilities as the difference between the scores of 98 and 99?

One of the essential preconditions to the standard rules of arithmetic is that one more unit should mean the same amount extra, no matter how much we already have.

In commenting about Stevens' definition with regard to the levels at which measurement occurs Bond and Fox (2007) refer also to the 'one more unit' problem by stating:

The interesting, but crucial difference ... is that while classification (referring to nominal) and seriation (referring to ordinal) are necessary precursors to the development of measurement systems they are not sufficient for measurement. The distinctive attribute of a measuring system is the requirement for an arbitrary unit of differences that can be iterated between successive lengths. (p. 4)

The problem of unequal units was first noticed by Thorndike (1904) who observed that even if one attempts to measure as simple a thing as spelling ability there exist no units in which to measure. If a list of words is arbitrarily constructed and the number spelled correctly is used to indicate ability one is struck by the inequality of units. 'All results based on the equality of any one word with any other are necessarily inaccurate' (Thorndike, 1904, p.7).

The Institute for Objective Measurement (2000) in order to emphasise the importance of one extra unit meaning the same amount throughout the construct continuum has given the following definition of objective measurement: "Objective measurement is the repetition of a unit amount that maintains its size, within allowable error, no matter which instrument, intended to measure the variable of interest, is used and no matter who or what relevant person or thing is measured".

Thorndike, 'the patriarch of educational measurement' (Wright, 1999, p.4) realised the unavoidable ambiguity in counting concrete events, however indicative they may seem. He was not only aware of the irregularity of the units counted but also of the non-linearity of raw scores.

Wright (1997, 1999) explains with the help of diagrams why raw scores are not linear.

The linear measures we intend raw scores to imply have no such bounds (referring to zero and maximum scores). Therefore a reasonable step from concrete counting to abstract measuring is required. (Wright, 1999, p.4)

The development of units of measurement which are arbitrary but can be iterated along a scale of interest so that the unit values remain the same has been the primary focus of Rasch measurement.

Rasch Measurement (Wright, 1989, Wright and Masters, 1982)

Rasch measurement begins with the idea of an attribute or a variable or a line along which objects can be positioned and the intention to mark off this line in equal units so that distances between points on the line can be compared.

A person's measure is his estimated position on the line of the variable. The instruments of observation are usually questionnaire and test items. The corresponding measure of an item (its calibration) is its estimated position on the line of the variable along which persons are positioned.

Persons are measured and items are calibrated on the variable which they work together to define. However, because items are accessible to invention and manipulation in a way that persons are not, it is useful to think of a variable as brought out (or defined) by its items.

The measurement of any object describes only one attribute of the object being measured. Further, only those characteristics of an object that can be described in terms of "more" or "less" can be measured (those characteristics that can be thought of as linear magnitudes).

In other words, the measurement of an object is in effect the allocation of the object to a point on an abstract continuum. If, for example, several people are described as to their weight, each person is allocated a point on an abstract continuum of weight.

Therefore, measurement implies the reduction or restatement of the attribute measured to an abstract linear form.

The basic requirements of measuring are:

- The reduction of experiences to a one-dimensional abstraction
- More or less comparisons among persons and items

- The idea of linear magnitude inherent in positioning objects along a line
- A unit determined by a process which can be repeated without modification over the range of the variable.

Underlying the idea of a variable is the intention to think in terms of more or less, that is the intention of order. The idea of order provides the basic ingredients from which measures are made.

A measurement model which will handle observations in a way that the relative strengths of persons and items can be compared along the variable must:

- Absorb the inevitable irregularities and uncertainties of experience.

The uncertainties of experience are handled by expressing the model of how person and item parameters combine to produce observable events as a probability. We do not try to specify exactly what will happen. Instead, we specify the probability of an indicative event occurring. This leaves room for the uncertainty of experience without abandoning the construction of order.

- Preserve the idea of order in the structure of the observations.

The idea of order is maintained by formulating measurement models so that the probabilities of success define a joint order of persons and items. The strongest of any pair of persons is always expected to do better on any item and the weakest of any pair of items is always expected to be done better by any person.

- Enable the independent estimation of distances between pair of items and any pair of persons by keeping item and person parameters accessible to sufficient estimation and inferential separation.

The measurement model must connect the observations and the person and item parameters in a way which permits any selection of relevant observations to estimate useful values for the parameters (i.e. measure must have generality). This can be done

effectively only when the formulation relates the parameters so that person parameters can be conditioned out of the model when items are calibrated to obtain sample-free item calibration, and item parameters can be conditioned out when persons are measured to construct test-free person measures.

The Rasch model is the only Item Response Theory model devised and successfully used so far that meets the above requirements.

Appropriateness Measurement (AM)

The branch of measurement which is concerned with the investigation of the inevitable irregularities contained in the data is called *Appropriateness measurement (AM)*. These irregularities are the unusual, aberrant or inappropriate individual score patterns. An aberrant score pattern is one that is improbable, given either that an IRT model fits the data or given the item score patterns of other persons in the group. Drasgow, Levine and Williams (1985) define AM as “a model-based attempt to control test pathologies by recognizing unusual patterns”.

Many researchers (such as Athanasou & Lamprianou, 2002; Harnish and Linn, 1981; Karabatsos, 2000; Linacre & Wright, 1994; Meijer, 1996; Molenaar & Hoijtink, 1996; Petridou and Williams, 2007; Rudner, 1983) have suggested various possible reasons leading to these unusual patterns.

These reasons include:

- cheating
- copying
- guessing
- carelessness
- extreme creativity
- alignment errors
- item multidimensionality
- misworded items
- distraction (one factor which may lead to distraction is Attention Deficit hyperactivity Disorder, ADHD)

- test anxiety
- special knowledge
- low language fluency
- class effect (non-standard administration practices, class cheating and instructional effects)

The validity of ability estimates of respondents with aberrant response patterns is questioned in the literature by many authors (such as Petridou and Williams, 2007; Reise and Flannery, 1996; Rudner, 1983; Smith, 1990; Wright and Masters 1982) but not thoroughly investigated.

Linacre and Wright (1994), Molenaar & Hoijtink (1996), Athanasou and Lamprianou (2002) are a few of the authors who suggest deeper investigation into the reasons behind aberrant response patterns, through interviews.

The main purpose of this study is to investigate reasons behind aberrant response patterns in a specific form of assessment, the classroom maths tests.

Performance Assessment and classroom tests

Assessment is generally the process of documenting, usually in measurable terms, knowledge, skills, attitudes and beliefs.

Academic performance assessment requires students to demonstrate that they have mastered specific skills and competencies by performing complex tasks or producing some work. It evaluates thinking skills such as analysis, synthesis, evaluation and interpretation of tasks, facts and ideas, skills which standardised tests generally avoid.

Mehrens (1992) states three influences that he believes contribute to the support for performance assessment:

- Selected-response tests usually, but not always, call only for recognition. Such tests fail to trace and mark higher-order thinking skills such as whether students can solve problems, synthesise or think independently.
- Cognitive psychologists believe that students should acquire both content and procedural knowledge. Particular types of procedural knowledge are not

accessible through selected-response tests. Therefore there is a call for an increased use of performance assessment in education.

- High-stakes tests will most likely continue to influence what teachers teach but performance assessment contribute to more worthy instructional targets than high-stakes tests.

Educators use performance tests to determine a student's status with respect to significant skills. Based on the student's level of achievement on the performance test the teacher makes an inference about the degree to which the student has mastered the skills that the test represents.

Classroom tests are performance tests that aim to measure learning outcomes that are specific to an in-depth study of the complex principles and skills related to the content material under consideration. Unlike standardised tests, classroom tests include a variety of item types including short-answer, constructed-response and performance tasks in addition to the traditional multiple-choice test questions.

In classroom maths tests for high school students, like the tests used in this study, it is common practice to include complex multistep problems which are designed to assess students' abilities to identify an appropriate solution strategy and to pursue it to a successful completion. Assessing all the steps in a student's performance on such problems gives more detailed information about the degree of mastery of the required skills and a more precise estimate of the students' abilities.

1.3 This study

In the vast majority of the literature on aberrant response patterns dichotomous data from high stakes tests (usually standardised tests), have been the main concern. Classroom achievement tests, and in particular maths tests, which are not usually high-stakes or dichotomously-scored tests have not been satisfactorily dealt with in studies with fit indices.

This project was not designed to be an evaluation of IRT models or of two misfit indicators (infit and outfit mean square statistics). Rather it was designed to use these two readily available and widely used indices to identify students with aberrant response patterns in classroom maths tests and to address the following research questions.

1. Which, if any, of the following factors that could lead to unexpected responses in classroom maths tests affect students' responses leading to aberrant response patterns?
 - Different schools.
 - Different teachers.
 - Student Gender.
 - Teacher Gender
 - Language competency.
 - Interest in mathematics.
 - Private tuition in mathematics.
 - Ability.
 - Test anxiety
 - Attention Deficit Hyperactivity Disorder (ADHD)
 - Maths Self esteem
 - Atypical schooling.
 - Item order.
 - Study time.

Although some of the factors contained in this list are the same as the ones mentioned earlier, this list differs from the one given earlier in that it contains all the factors which

have been investigated through statistical techniques. Other reasons have been investigated through interviews.

2. Are there any other reasons that lead students to unexpected responses?
3. Do the same students consistently misfit over administrations of different maths tests? In other words is misfit an inherent characteristic of some students?
4. Are the predictive validity and reliability (internal consistency) of scores of misfitting students of a lower degree than scores of fitting students as suggested in the literature?
5. How are the infit and outfit mean square statistics affected by unexpected responses?
 - a. How much does one unexpected response contribute to the categorization of a response pattern as misfitting through the outfit mean square statistic?
 - b. How many well-targeted 'less likely' responses are needed to categorise, through the infit mean square statistic, a response pattern as aberrant?

The word "factor" is used in this study for all demographic or psychological characteristics that were considered. The possible associations of these factors with misfit were investigated through statistical methods.

On the other hand the word "reasons" was used for what students themselves give as explanations for their unexpected responses.

For the purposes of this study data was collected over two academic years (2004 – 05 and 2005 – 06) from first form students, of age around 15, from 5 lyceums in 2004 – 05 and 3 lyceums in 2005 – 06 in Cyprus.

The educational system in Cyprus

In the Educational system in Cyprus, children attend the primary school for six years (ages 6-12). For the academic year 2007 – 2008 there are 349 primary schools in Cyprus attended by about 52,500 pupils.

After primary, comes the secondary education which is divided into two phases. First, pupils have to attend a Gymnasium, for 3 years (ages 12-15) and then they have a choice of two different directions:

- The Lyceums, which are attended by the vast majority of the gymnasium leavers, and usually the more academically gifted.
- The Technical Schools, attended by students inclined more towards technical or hotel oriented professions.

Overall, there are 76 Gymnasias, 44 Lyceums and 11 Technical schools in Cyprus.

There are about 28,650 students in the gymnasias, 24,300 in the lyceums and 4,500 in the technical schools. (About 84% of the students who finish the gymnasium continue in the lyceums whereas the remainder in the technical schools)

Originality of the study

The originality and importance of this study, compared with the bulk of the research in the subject, lies in the fact that low-stakes, maths classroom tests are used.

Classroom tests are by far the most widely used form of testing in the world. The researcher found that in his school, 80 maths tests were administered during the first term of the academic year 2007 – 2008 (the first term had 65 school working days). From this number and the number of schools in Cyprus the researcher made a rough estimate of the number of maths tests administered to the whole student population (primary and secondary) in Cyprus over the first term. The estimate was 37,600 tests for the first term which gives on average of about 580 tests per day. This number is just for maths tests per day, just in Cyprus.

From this estimate one can realize that most probably quite a few million tests are used every day in the world in the classroom setting, leading to ability estimates of enormous numbers of students.

This familiar setting, the classroom setting, with the intimacy between assessor (classroom teacher) and student, and the not so high importance placed on the results, in terms of decision making about the future of the students (unlike some high-stakes tests) makes this kind of testing a low-stakes event. In this testing situation perhaps factors like test anxiety play a reduced role in affecting students towards unexpected responses.

Also multistep problems are used with partial credit awarding, which although not thoroughly explored in the literature, give more detailed information about the skills acquired by the students and consequently more accurate estimates of students' abilities.

For the analyses of this kind of data the Partial Credit Rasch Model was used as opposed to the much more commonly found in the literature Dichotomous Rasch Model.

Finally, interviews of highly misfitting students were conducted in an attempt to investigate further the reasons, as perceived by the students themselves, for unexpected responses. Such an investigation was reported only in one study in the literature (Petridou and Williams, 2007).

CHAPTER 2: REVIEW OF THE LITERATURE

This chapter provides a description of educational and psychological tests and addresses validity and reliability issues. It also describes the two major testing theories, Classical Test Theory (CTT) and Item Response Theory (IRT), introduces the Rasch models and appropriateness measurement, focusing on the various person fit statistics which are used to identify misfitting students.

Finally the infit and outfit mean square statistics and their critical values are described in detail discussing also criticisms against them.

2.1 Tests in general

2.1.1 Introduction

The Standards for educational and psychological testing give the following definition for a test:

A test is an evaluative device or procedure in which a sample of an examinee's behavior in a specified domain is obtained and subsequently evaluated and scored using a standardized process. (AERA, APA and NCME, 1999, p. 3)

In general, a test, educational or psychological, implies a presentation of a set of questions to be answered in order to obtain a measure of a characteristic of a person.

The way test scores are interpreted categorizes tests into two types, norm- and criterion-referenced tests.

In norm-referenced tests, a score is interpreted by comparing it with that of a large group of individuals, called the norm group. The emphasis in such tests is on what position amongst the norm group a person holds based on his/her score. Mehrens and Lehmann (1991, pp.19 – 20) give a detailed description of uses of the norm-referenced

tests, which include differential prediction in aptitude tests, decision making for vocational or educational planning and in selection decisions.

In criterion-referenced tests, score interpretations are made by comparing the score with some specified behavioral domain, or criterion of proficiency.

Lyman (1998) states:

As adopted in general school use today, the criterion-referenced test is typically one of a series of coordinated achievement tests that is designed to measure a single behavioral objective within a course of study. . . . In practice, the teacher strives for pupil mastery of the material. . . . A (criterion-referenced) test is used to evaluate pupil mastery of each unit. (p. 33)

Criterion-referenced tests are very important in education and particularly in classroom assessment. They can be used in mastery testing, minimum competency testing, licensure testing and for instructional decisions within the classroom.

2.1.2 Classroom tests

The classroom achievement test is made from a set of items administered to pupils through which the teacher can (hopefully) reliably and validly evaluate how effectively his or her students have learned what has been taught. They are assessment tools that help the teachers with one or more of the following (as discussed by Mehrens and Lehmann, 1991):

- evaluating a student's overall achievement and growth in a content domain
- assigning grades to students
- improving their teaching methods
- ascertaining the effectiveness of the curriculum
- diagnosing students' weaknesses and providing feedback to them and remedial instruction
- diagnosing students' strengths and providing enriching work
- encouraging good study habits
- planning review materials

- identifying potential issues to be faced
- deciding about grouping of pupils in a class
- determining the pace of instruction in the classroom and
- reporting achievement to parents.

Popham (2000) places emphasis on the contribution of tests to 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.

Also Rudman (1989) argues that teachers tend to use tests that they prepared themselves much more often than any other type of test to monitor what has been previously learned.

Commenting on the importance of classroom tests Mehrens and Lehmann (1991) state:

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 (p. 79).

Comparing Teacher-made and standardized achievement tests

Standardized tests are commercially prepared measuring instruments for which the authors carefully delineate the administrative and scoring procedures. Scoring is usually objective although essays and other open-ended items may be included in the test. The standardized test is usually administered to a norm group first so that any person's performance can be interpreted in a norm-referenced manner.

These two types of test are more alike than it might first seem since the objective of both is to measure pupil knowledge, skills and ability. Any test that has a representative sampling of the relevant content and that is designed to measure the extent of present knowledge and skills is an achievement test, regardless of whether it was constructed by a classroom teacher or by a professional test-maker.

Mehrens and Lehmann (1991, pp. 346-350) give a detailed account of the differences between standardized and teacher-made tests with respect to the following aspects:

Sampling of content

Standardized tests are traditionally designed to cover more than one year's learning whereas teacher-made tests usually cover a single unit of work or that of a term. Therefore the standardized test covers much more material.

Construction

The two types of test differ in the relative amount of time, money, effort and resources that are available for their construction.

According to Mehrens and Lehmann (1991), the following steps are common in the procedure for constructing a standardized test:

- The test publisher arranges a meeting of curriculum and subject matter experts who will study thoroughly the syllabi, textbooks and programs throughout the country.
- A list of objectives is prepared (information pupils should have, principles they should understand and skills they should possess).
- A table of specifications is outlined that will guide the test-makers in constructing the test.
- With the assistance of classroom teachers and subject matter experts a team of professional test writers prepares the items.
- Instructions to both administrators and pupils are written.
- Tryout tests are given to a sample of pupils for whom the test is designed.
- Item analysis is carried out to identify poor items.
- Comments from the test administrators pertaining to timing and clarity of instruction are noted.
- The test is ready to be standardized. The refined test is administered to a representative sample of pupils and scored.
- Reliability and validity evidence is obtained.
- Norms are prepared for the standardization sample.

In classroom tests however, the teacher alone constructs the test and usually has a limited amount of time to devote to test construction. He or she often does not have the time to examine the items in terms of difficulty and discrimination, or to try out his test beforehand in order to clarify any ambiguous directions or to alter the speededness of the test by adding or removing items.

Ideally other teachers should review every classroom test critically to minimize any deficiencies.

Classroom teachers should not develop an inferiority complex because of these remarks. "They should recognize that they have been trained to be teachers and not test-makers" (Mehrens and Lehmann, 1991, p. 349)

Reliability

Standardized tests generally have high reliability, often over 0.90, and small standard errors of measurement whereas, the teacher-made tests' reliability is generally unknown although, if carefully designed the reliability can be high.

Interpretive aids

Standardized tests usually provide material accompanying the test with suggestions for teaching or reteaching the concepts pupils do not understand.

Norms

Standardized tests provide norms. With national norms, one can make numerous comparisons of the performance of individual students, classes, grades, schools and school districts.

Teacher made tests do not have norms, or if they do have these will be at best locally based.

Purposes and use

Standardized tests are usually constructed to measure generally accepted objectives. They have a broad sampling of content and can be used to measure the general level of achievement of pupils and may be too general to meet the objectives of a particular school or teacher at particular times.

Teacher-made tests, on the other hand, will have narrow content sampling but usually measure more adequately the degree to which the objectives of a particular course for a particular teacher have been met. In other words they can assess specific classroom objectives more satisfactorily than standardized achievement tests.

Teacher-made tests can be more useful than commercially made standardized tests because they are more closely related to a teacher's particular objectives. Given also the amount of time and effort needed to construct standardized tests, teacher made tests are more flexible and adaptive to curricula changes,

Because the standardized and teacher-made achievement tests serve different purposes, school personnel should consider the supplemental value of standardized achievement test scores to teacher-made test scores and teacher observations and judgments, rather than argue that one measurement device is better than the other.

(Mehrens and Lehmann, 1991, p. 349)

2.1.3 Validity

Validity is the most important consideration in test evaluation.

Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores. ... What is evaluated is not the test but the inferences derived from the test.
(Messick 1993, p.13)

Experts agree that validity (or construct validity) is a unified concept.

The standards (1999) specifically state that:

Validity is a unitary concept. It is the degree to which accumulated evidence supports the intended interpretation of test scores for the proposed purpose. (AERA et al., 1999, p. 11)

Conventional view of validity

In a validation study it is useful to use three different categories, *construct validation*, *criterion validation* and *content validation*. The different category labels used are by no means distinct types of validation, they are just facets of the same unitary concept.

The evidence gathered for each of the categories is different; however, when the results of the studies are put together they provide an assessment of the overall validity of the test.

Construct-Related Validation

A construct is a variable, which is abstract and latent rather than concrete and observable. Such a variable is literally something that scientists 'construct' (put together from their own imaginations) and which does not exist as an observable dimension or behavior.

(Nunnally and Bernstein, 1994, p.85)

Cronbach and Meehl (1955) state that construct validation is involved whenever a test is interpreted as a measure of some attribute or quality that is not operationally defined. The problem that has to be solved in such a validation is what constructs account for the variance in the test performance.

Construct validation is linked to the theoretical basis of the construct.

Sources of evidence for the construct interpretation include:

- Intercorrelations between the responses to the items, tasks or parts of the test may be used to support the assertion that a test measures primarily a single construct. (Factor Analysis is commonly used for this purpose)
- Substantial relationship of the test scores to other measures of the same construct.
- Absence of relationships of the test scores with measures of different constructs.
- Investigations of differences in these relationships and structure
 - over time
 - across groups or settings

- in response to experimental interventions (such as instructional or therapeutic treatment or motivational conditions).

“These varieties of evidence are not alternatives but rather supplements to one another. This is the main reason why validity is now recognized as a unitary concept”. (Messick, 1993, p. 16)

Criterion-Related Validation

The investigator is interested in some criterion he or she wishes to predict. He or she administers the test and then computes a correlation of the test scores with an independent measure of the criterion. This type of validation evidence can also be used in construct validation, since it gives further support to the hypothesis that the construct measured is the intended one.

If the criterion is obtained some time after the test is given, the investigator is studying *predictive validity*. If the test scores and the criterion scores are obtained about the same time then he or she is studying *concurrent validity*.

The ‘criterion problem’ however is what to measure, how to measure it and whether this measurement is free from bias.

Another problem of concern in this type of validity study is how accurately criterion performance can be predicted from scores on a test. If the test under investigation relates to school based achievement then the criteria could include: aptitude test scores, grade point average or supervisor’s ratings.

Content-Related Validation

A content validation study differs from the other two types in the sense that, as stated by Nunnally and Bernstein (1994, p 84) “construct and predictive validity usually stress correlations among various measures, but content validation is largely based upon the opinions of various users.”

Content related evidence takes the form of consensual professional judgments about the relevance of item content to the specified domain and about the representativeness with which the test content covers the domain content. A test is representative if it

reproduces the essential characteristics of the universe in their proper proportion or balance.

Messick (1993) comments that the major problem with content validity is that “it is focused upon test form rather than test scores, upon instruments rather than measurements” (p. 41).

From this point of view it is logical to question the appropriateness of a content validation study when assessing validity. However Messick (1993), in order to emphasize the importance of such a study states that “in the fundamental sense so-called content validity does not qualify as validity at all, although such considerations of content relevance and representativeness clearly do and should influence the nature of score inferences supported by other evidence” (p. 17).

Therefore, according to Messick, although content validity may not qualify as validity per se, the test users should take into account the relevance and representativeness of the test content, together with other appropriate sources of evidence of construct validity, before making their inferences.

According to Mehrens and Lehmann (1991) a content validation entails the following steps:

- Defining the performance domain of interest
- Selecting a panel of qualified experts in the context domain
- Ranking or weighting the objectives in terms of their importance before matching items to objectives.
- Every element of the assessment instrument being judged by the experts on its relevance, representativeness and clarity.
- Collecting and summarizing the data from the matching process.

Such a validation procedure is commonly used in evaluating achievement tests, which are designed to measure how well an individual has mastered specific skills or course of study.

Teacher-made (or classroom) tests fall in this category of tests and therefore content validity is of major importance to the validation of such tests.

2.1.4 Messick's modified view of validity

Messick (1993) writes:

The continuing enumeration of three categories of validity evidence perpetuates ... the temptation to rely on only one (or, worse still, any one) category of evidence as sufficient for the validity of a particular test use. (p. 20).

Furthermore, the conventional view is incomplete because it fails to take into account evidence of the value implications of score meaning as basis for action as well as the social consequences of score use.

Messick (1993, 1995) reemphasizes the fact that validity is a unified and many-faceted concept. Referring to validity as a unified concept does not necessarily imply that it cannot be usefully differentiated into distinct aspects which will address issues that might otherwise be overlooked. "The intent of these distinctions is to provide a means of addressing functional aspects of validity that help disentangle some of the complexities inherent in appraising the appropriateness, meaningfulness, and usefulness of score inferences." (Messick, 1995, p. 5)

Messick (1995, pp 6-8) describes six distinguishable aspects which function as general validity criteria or standards for all educational and psychological measurement:

The content aspect of validity includes evidence of content relevance, representativeness and technical quality.

It is not sufficient to merely select tasks that are relevant to the construct domain. The assessment should include tasks that are representative of the domain in an effort to ensure that all important parts of the construct domain are covered. Both the representativeness and relevance of assessment tasks are traditionally appraised by expert professional judgments.

The substantive aspect emphasizes firstly the need for tasks providing appropriate sampling of domain processes in addition to the traditional coverage of domain content and secondly the need to go beyond professional judgments of content to collecting

empirical evidence that the intended sampled processes are actually engaged by the respondents in task performance.

The structural aspect appraises the extent to which the internal structure of the assessment reflected in the scores is consistent with the structure of the construct domain.

The selection or construction of the assessment tasks together with the rational development of scoring criteria and rubrics should be guided by the theory of the construct domain.

Thus, the internal structure of the assessment (i.e. intercorrelations among the scored aspects of task and subtask performance) should be consistent with what is known about the internal structure of the construct domain.

(Messick 1995, p. 7)

The generalizability aspect examines whether the score properties and interpretations can be generalized to and across population groups, settings and tasks.

Evidence of generalizability depends on the degree of correlation of the assessed tasks with other tasks representing the construct or aspects of the construct.

The external aspect includes convergent and discriminant correlations with external variables.

It refers to the extent to which the high or low relationships of the assessment scores with other measures and nonassessment behaviours reflect the expected relations implicit in the theory of the construct being assessed.

Thus, the meaning of the scores is substantiated externally by appraising the degree to which empirical relationships with other measures, or the lack thereof, is consistent with that meaning.

(Messick 1995, p. 7)

The consequential aspect refers to the social consequences of the score interpretations of the assessment.

It is important to accrue evidence of the positive expected consequences of the assessment, such as benefits for teaching and learning, as well as evidence that adverse consequences are minimal. This includes collecting evidence for evaluating the intended and unintended consequences, especially with regard to bias, or fairness in test use.

Of primary importance with respect to adverse consequences is that low scores should not occur because of construct underrepresentation (that is, the assessment missing something important to the focal construct that, if present, would have permitted students to display higher competence) or because of construct-irrelevant variance (that is, assessment containing something irrelevant that interferes with the affected students' demonstration of competence)

Messick (1995) concludes that these six aspects of construct validity apply to all educational and psychological measurement and they provide a way of addressing the multiple and interrelated validity questions that need to be answered in justifying test interpretation and use.

Perhaps one of the most important procedures employed in a validation is the study of intercorrelations between the responses to the test items in an attempt to provide evidence supporting the hypothesis that the test measures one construct. The statistical method used in such a study is called Factor Analysis.

2.1.5 Factor Analysis

Kline (1994) explains factor analysis in detail.

Factor analysis is a highly complex statistical procedure, which was used for the first time in 1904 by Spearman. It is a statistical method for simplifying complex sets of data, usually starting with correlation matrices.

Factor analysis gives a mathematical account of these correlations in terms of a few factors, which can easily be understood. A factor is a construct or dimension, which indicates the relationship between a set of variables and is operationally defined by its factor loadings. A factor loading is simply the correlation of a variable with the factor.

When we factor analyze the correlations between the items of a test, we obtain a set of factors on each of which the items load (correlate). Depending on which items load on which factor, we try to define the factors.

Squaring a correlation coefficient between two variables indicates how much common variance there is in the two variables. Therefore if we square and add all the factor loadings for each item, this gives the communality (h^2), which is, clearly, the total variance of the item, which the factors “explain”.

The size of a factor, large or small, is computed by averaging across items its squared factor loadings. This computation yields the percentage of variance accounted for by the factor. The raw sum of squares of the factor loadings is referred to as the eigenvalue of the factor.

The initial condensation

The first computation of factor analysis is condensation, which reduces the complexity of the correlation matrix by condensing the variables into factors. It can be done by different methods.

Principal components analysis (PCA) vs. Principal factor analysis (PFA)

The two methods are identical except that instead of unity in the diagonal of the correlation matrix, in PFA some other estimate of the communality is inserted. This means that while the PCA explains all the variance in the given matrix, thus incorporating error variance in the items into the factors, the PFA does not. Nunnally and Bernstein (1994) argue that if good measures of reliability are available and one is confident about the number of common factors underlying the data, reliability coefficients can be placed in the diagonal of the correlation matrix instead of unity and PFA performed.

PFA theoretically has an advantage, because it is unlikely that factors could “explain” all the variance in any given matrix and, since all correlations contain error, the full account of principal components must be contaminated by error. However, Kline (2000, p. 58) agrees with Harman (1976) that in large matrices the differences between PCA and PFA are negligible.

Maximum likelihood factor analysis (MLA)

MLA is a method which produces estimates of the population factors from the sample correlation matrix. The main advantage of MLA is that there is a statistical test for the number of factors, which is a problem with the other methods. However, for statistical reasons, large samples are required for this procedure. Kline (2000) suggests more than 1000. Also, in practice, with robust factors, MLA gives results identical to the other two methods.

Selecting the right number of factors

In the mathematics of factor analysis each variable (or item) is assumed to have an eigenvalue of one. Thus, a factor to be of any importance must have an eigenvalue greater than 1; otherwise it would account for less variance than an item and would be trivial both psychologically and statistically.

After the initial condensation and selection of factors, and before any interpretation can be made, factors may be rotated. Rotations make the interpretation easier.

Factor space is a multidimensional space, having as many dimensions as factors, where the axes represent the factors. In this space each item is plotted and the coordinates of the item directly map on to the factor loadings. The whole set of axes (factors) can then be rotated to any position.

By rotating the factors, the item loadings are changed, but the communalities are not. Indeed there is an infinite number of possible different solutions (rotations).

Kline (1994) explains that the different factor analytic solutions are mathematically equivalent in that they explain the same amount of variance in each variable (item). Furthermore the rotated factors reproduce the original correlations precisely as well as the unrotated solution. The formula for computing the correlations is:

$$r_{xy} = r_{x1y1} + r_{x2y2}$$

where r_{xy} is the correlation between variables x and y ;

r_{x1y1} is the cross product of the factor loadings of variables x and y on factor 1;

r_{x2y2} is the cross product of the factor loadings of variables x and y on factor 2.

In fact there are infinite mathematically equivalent rotations to a factor analysis. Which one do we choose then? According to Kline (2000, p. 59) Thurstone (1947) suggests that the simplest solution is the best. The simplest solution (or simple structure) is obtained when each factor has a few high loadings with the majority being zero.

When wishing to produce good factor analysis the following should be borne in mind:

- Sample size. A minimum of 100 subjects is required to avoid too much error in the correlation matrix.
- Subject to variable ratio. If there are more variables than subjects factor analysis is meaningless. With clear factors a ratio of 2:1 yields replicable results.
- Rotation to simple structure should be carried out by Varimax for orthogonal factors or Direct Oblimin for oblique (correlated) factors as best fits the data.

The Varimax method, which aims at simple structure while keeping the factor axes orthogonal (uncorrelated) and Direct Oblimin are methods for rotating factors.

Although in some instances simple structure cannot be obtained with orthogonal factors, where this is possible it is generally agreed that Varimax is the most efficient procedure. Varimax aims to maximize the sum of variances of squared loadings in the columns of the factor matrix. This produces in each column (which is, of course, a factor) loadings which are either high or near zero. This is one of the critical features of simple structure.

(Kline, 1994, p. 68)

Direct oblimin is suggested by Kline (1994, 2000) as the best amongst many methods for obtaining simple structure when one has oblique or correlated factors.

Exploratory and Confirmatory Factor Analysis

Exploratory factor analysis, the one described above, aims to explore the field, to discover the main constructs or dimensions in the data. Spearman, in 1904, originally developed factor analysis in the area of human abilities in order to answer the question: 'What constructs or dimensions could account for the correlations between abilities?' (Kline, 1994, p. 7).

Confirmatory factor analysis was developed much later (in 1973 by Joreskog).

In this method, based upon previous studies or on relevant theory, factor loadings for the variables are hypothesised. Confirmatory factor analysis then proceeds to fit these loadings in the target matrix, as it is called, as closely as possible. How good the fit is can also be measured. Since the scientific method, ... involves testing hypotheses, confirmatory analysis has become acceptable to psychologists who were previously resistant to exploratory methods. ... in the social sciences it is often so difficult to specify with any precision what the factor loadings should be that confirmatory analysis is not highly useful. (Kline, 1994, pp. 10 – 11)

Objections to factor analysis

Kline (1994, pp. 11 – 12) discusses some objections to factor analysis giving his responses to them. These include:

1. The main objection is that there are an infinite number of mathematically equivalent solutions. This is true; however psychometricians have developed powerful methods for choosing the right solution.
2. Factor analysts often disagree as to what are the most important factors in the field. This often results due to poor factor analytic methods.
3. It is difficult to replicate factor analyses. This stems from the first objection and with sound methodology it can be overcome.
4. It is sometimes said that with factor analysis you only get what you put in so it is difficult to see how the method can be useful. This objection is sometimes valid. For example, if in a study of abilities no measures of musical ability were included then no factor of musical ability could emerge. That is why in

exploratory analyses it is essential to sample variables as widely as possible. 'However, generally this is not so and, ironically, one of the most attractive aspects of factor analysis as a statistical method is that it can reveal constructs which ere previously unknown.' (Kline, 1994, p.12)

One of the most important uses of factor analysis is perhaps its use as a powerful tool in assessing the dimensionality of test data in construct validation studies.

2.1.6 Reliability

Whenever a test is administered various sources of error cause variation in a person's score. These sources include:

- Trait instability (the characteristic being measured may change over time)
- Sampling error (the particular questions asked to infer a person's knowledge)
- Administrator error (changes in directions, timing or rapport with the test administrator)
- Scoring error (inaccuracies in scoring the test)
- Things like motivation, concentration, fatigue and health, good or bad luck.

(Mehrens and Lehmann, 1991)

Reliability can be defined as the degree of consistency or reproducibility of test scores. It is theoretically defined as the proportion of variation in the observed scores attributable to the variation in the true scores. Reliability is a necessary but insufficient condition for valid score-based inferences.

Classical Test Theory starts with the model, $X = T + e$, where X is the observed score of an examinee on the test, T the true score (which is conceptualized as the hypothetical average score resulting from many repetitions of the test or alternate forms of the instrument) and e the error.

The model has the following assumptions:

- (1) T is constant, changes in X are due to error
- (2) Errors are random and they do not correlate with T or with each other.

These assumptions together with the theoretical definition that: reliability is the proportion of variation in observed scores attributable to the variation in the true scores (i.e. r_{xx} = variance of true scores/variance of observed scores) have led to the following formulae about the reliability and the standard error of measurement:

Reliability (r_{xx}) and standard error of measurement (SEM)

$r_{xx} = 1 - \frac{S_{\varepsilon}^2}{S_x^2}$ scores and $SEM = S_x \sqrt{1 - r_{xx}}$	where	$\left\{ \begin{array}{l} S_x^2 = \text{Variance of group's observed} \\ S_{\varepsilon}^2 = \text{Error variance} \\ SEM = \text{Standard error of measurement} \end{array} \right.$

Just as the total group has a standard deviation, theoretically each examinee's personal distribution of possible observed scores around the examinee's true score has a standard deviation. When these individual error standard deviations are averaged for the group, the result is called **standard error of measurement (SEM)**

(Crocker and Algina, 1986, p. 122)

If we accept the premises we can be 68% confident that the true score of an examinee lies in the interval $[X - 1 \text{ SEM}, X + 1 \text{ SEM}]$ and 95% confident that it lies in the interval $[X - 2 \text{ SEM}, X + 2 \text{ SEM}]$.

(Mehrens and Lehmann, 1991) and (Crocker and Algina, 1986).

In general when educators think about the reliability of a test they focus on the consistency with which the test is measuring whatever is measuring. Different approaches to test reliability yield three substantially different ways of viewing this consistency.

Popham (2000) describes in detail these different ways as follows:

Stability reliability

Stability estimates of reliability are based on the consistency of a test's measurement over time. In this case:

Reliability coefficient = test – retest correlation coefficient.

The time interval between the two testing occasions is however crucial. It must be selected so as to reduce the influence of the first testing on the second, but at the same time to reduce the likelihood of events in the life of the students distorting the second set of test results.

Popham (2000) suggests a time interval of a few weeks.

This sort of information is not easy to obtain in the classroom setting where as soon as the test is administered and marked weaknesses are identified and remedial work is suggested to help students overcome those weaknesses. It is therefore, not always possible to readminister the same test after explanations about the test items have been given.

Alternate-form reliability

Alternate-form reliability refers to the consistency of measured results yielded by different forms of the same test.

For this, content-parallel tests are needed. However, assertions about content similarity are not sufficient. Correlational evidence, students' means and standard deviations on the two forms are also required.

Internal consistency reliability

In internal consistency reliability the focus is in the homogeneity of the set of items that make up the tests. That is, whether all the items function in a similar fashion.

Calculating estimates of internal consistency reliability (Traub, 1994, 75-95)

1. The Split-Half Method

The correlation (r_{yy}) between the scores, on parallel half-tests provides the estimate of the reliability of either half test. The reliability r_{xx} of the full-length test is then estimated by using the Spearman-Brown Formula.

$$r_{xx} = \frac{2r_{yy}}{1 + r_{yy}}$$

The drawback in this method is the difficulty of assuming that the half-tests are parallel.

2. Rulon's Formula and non-parallel test components.

Without the assumption of parallel half-tests, Rulon's formula gives a lower bound to reliability.

$$r_{xx}^2 = 2 \left[1 - \frac{\sigma_{y1}^2 + \sigma_{y2}^2}{\sigma_x^2} \right]$$

where: σ_{y1}^2 and σ_{y2}^2 are the variances of the observed scores on the two parts and σ_x^2 is the variance of the observed scores on the full length test.

3. Kuder-Richardson Formula 20

The KR_{20} formula is focused on tests composed of dichotomously scored items.

According to Traub (1994), Novick and Lewis (1967) state that, if it is impossible to assume that the standard errors of measurement for an examinee on the different parts are neither equal nor necessarily related in a simple way, then this formula estimates a lower bound to the reliability of the test (KR_{20}).

$$KR_{20} = \frac{n}{n-1} \left[1 - \frac{\sum_{i=1}^n \rho_i(1-\rho_i)}{\sigma_x^2} \right]$$

where KR_{20} is the estimate of the reliability,

ρ_i is the proportion who answer the i th item correctly.

σ_x^2 is the variance of the observed scores on the full length test.

n = number of items in the test.

4. Cronbach's Alpha: (Introduced by L.J. Cronbach in 1951).

This coefficient is a generalization of the KR20 formula to apply also to tests where the items are not scored dichotomously. It is very useful when a test is composed of items on which the examinees' scores can take any value on a continuous scale. Alpha is considered as the lower bound to a theoretical reliability coefficient and is given by:

$$\alpha = \frac{n}{n-1} \left(1 - \frac{\sum_i \sigma_{yi}^2}{\sigma_x^2} \right)$$

where α = Estimate of reliability, n = number of items in the test

σ_{yi}^2 = Variance of the observed-score random variable for the i^{th} item

σ_x^2 = Variance of the observed-score random variable for the Total Score

Alpha is preferable over the other internal consistency estimates for two reasons. First, it can be used for both dichotomously and polytomously scored items. Therefore it can be used for tests with multiple-choice, true-false, Likert-scaled, constructed-response and essay-type items. Second, alpha requires only one test administration to be estimated, like the split-half coefficient. However, the split-half coefficient has the drawback that it is determined by how one groups the items. Alpha, on the other hand, is the mean of all possible split-half coefficients.

Standard error and confidence interval for alpha

Cronbach's alpha is by far the most commonly used index of internal consistency. A common research scenario that would benefit from reporting the ASE in conjunction with the coefficient alpha is the assessment of rivaling tests measuring the same construct. If the tests possess comparable alpha reliabilities, the ASE will provide evidence of the superiority of one over the other. A second scenario, and perhaps a more interesting one from the researcher's point of view, is when a testing organization is trying to refute claims of bias against a subpopulation. If the organization wished to demonstrate that the strength of the relationship between the test and some criterion was no different from that relationship in another population, the reliabilities, as well as their high and low estimates (confidence interval) would need to be considered.

Iacobucci and Duhachek (2003) and Duhachek and Iacobucci (2004) used the asymptotic distribution for the maximum likelihood estimator (MLE) of the variance of coefficient alpha (derived by Zyl, Neudecker and Nel, 2000), based on the standard statistical assumption of multivariate normality, to present the estimate of alpha's standard error (ASE) and consequent confidence interval.

Duhachek and Iacobucci (2004) compared their ASE and confidence intervals with alternative methods for computing confidence intervals. They concluded that their estimate, together with Feldt (1965) and Hakstian and Whalen (1976) were more precise than other methods.

Iacobucci and Duhachek (2003) investigated the effects of the number of items, the item intercorrelations and the sample size on the confidence intervals and concluded that:

- The confidence intervals are tighter (more precise estimation of alpha) as the item correlations increase.
- The confidence interval is always wider for smaller sizes, although as n increases ($n > 100$ in this case) and number of items increase ($p > 7$) no significant differences arise, given that the average item correlation $\bar{r} > 0.4$.
- The effect of sample size is the case of gaining power as one obtains more information. However, a sample of size 200 is not much more effective in obtaining precise estimates than a smaller sample ($n = 30$) if p and/or $\bar{r}$ is large.

Koning and Franses (2003) simplify the formula for the Iacobucci and Duhachek (ID) confidence intervals by using Zyl et. al (2000) result which states that if the items are parallel and $n \rightarrow \infty$ then the variance of alpha can be estimated by $\hat{Q} = \frac{2k}{k-1} \cdot (1 - \hat{a})^2$, where k is the number of items.

Koning and Franses (2003) then introduce two more methods for estimating confidence intervals for alpha, one asymptotic (involving again the standard normal and $\hat{Q}$) and one exact (involving the F distribution). The three methods are shown in table 2.2.

Table 2.2. Methods for estimating confidence intervals for alpha mentioned in Koning and Franses (2003).

Source	95% confidence interval
Iacobucci and Duhachek	$\hat{a} \pm 1.96 \cdot (1 - \hat{a}) \sqrt{\frac{2k}{n(k-1)}}$
(Asymptotic bounds)	(where n = sample size and k = number of items)
Koning & Franses	
(asymptotic bounds)	$1 - (1 - \hat{a}) \exp \left[\pm 1.96 \sqrt{\frac{2k}{n(k-1)}} \right]$
Koning & Franses	
(exact bounds)	$a_L = 1 - \frac{1-R}{F_L} \quad \text{and} \quad a_R = 1 - \frac{1-R}{F_R}$
	(where F_L and F_R are values of the F-distribution with $n(k-1)$ and n degrees of freedom such that $P(F < F_L) = P(F > F_R) = 0.025$)

Koning and Franses (2003) compared the ID confidence intervals with the two they proposed arguing that their exact confidence interval had a simulated nominal coverage approximately equal to the confidence level of 0.95. They concluded, however, that for large values of k and n the differences between the methods get smaller. In fact, in their study (where k took the values of 2, 4 and 6 only and n the values 50, 100 and 200) one

can see that the largest differences occur when $k = 2$ (an unrealistic number of items for a test) and $n = 50$. As $k \rightarrow 6$ and $n \rightarrow 200$ no differences existed between the nominal coverage of the ID and the exact confidence intervals.

Therefore, Koning and Franses (2003) concluded that as k increases there are no real differences between the precision of the three confidence interval estimates.

Desirable values of the reliability coefficient

Factors like, test length, group homogeneity, difficulty and objectivity can influence reliability.

With regard to the desirable values of the reliability coefficient Nunnally and Bernstein (1994) suggest the following:

If important decisions are made with respect to specific test scores, a reliability of 0.90 is the bare minimum and a reliability of 0.95 should be considered the desirable standard. However, never switch to a less valid measure simply because it is more reliable. (p.265)

Uses of the reliability coefficient

The reliability coefficient, together with the observed-score standard deviation, can be used to obtain an estimate of the standard error of measurement which

- Can then be used to calculate a confidence interval for the test taker's true score.
- Provides an impression of the variability that would be expected in a person's observed scores.

The reliability coefficient can also be used to compare the relative merits of two or more instruments being considered for the same application.

2.1.7 Item Difficulty and Discrimination in Classical Test Theory

The basis of classical test theory was described in section 2.1.6 on reliability.

In **item analysis**, psychometricians use two basic measures, item difficulty and item discrimination.

Item difficulty

This index is calculated by dividing the mean score of the item by the maximum possible score.

If items have only one correct answer, which is worth one point, then this index represents the percentage of examinees responding correctly.

Item difficulty clearly depends on the ability of the group of test takers. This affects also the distribution of scores. In high ability groups the distribution is negatively skewed whereas in low ability groups it is positively skewed. It is preferable to add/revise or delete items so that the score distribution in the target group is approximately Normal.

(Anastasi and Urbina, 1997, 177-178).

Item Discrimination (D)

To estimate D the test papers are arranged in order, based on the total score. Then two groups are identified, the high scorers and the low scorers. According to Crocker and Algina (1986) a classic study by Kelley in 1939 demonstrated that a more sensitive and stable item discrimination index can be obtained by using the upper 27% of the papers and the lower 27% (than by using the top 50 % and lowest 50 % suggested, mainly for small samples). "However, when sample size is reasonably large, virtually the same results can be obtained with the upper and lower 30% or 50%" (Crocker and Algina, 1986, p.314)

D for any item is then the difference of the average scores of the two groups for the specific item, divided by the maximum possible score on the item.

Interpreting the index of discrimination (D). Crocker and Algina (1986) propose the following interpretations for various values of D.

If	$D \geq 0.40$	the item is functioning satisfactorily
	$0.30 \leq D \leq 0.39$	little or no revision required
	$0.20 \leq D \leq 0.29$	item is marginal and needs revision
	$D \leq 0.19$	item eliminated or revised.

(Crocker and Algina, 1986, p.315)

However these ranges are not really set in stone. They can be used as indications of possible revision of certain items rather than for discarding them.

For example, low discrimination could mean that the item is too easy or too difficult. However it could be deliberately too easy for encouragement and motivation purposes. Such an item should not be removed.

Correlational indices of discrimination

The higher the correlation between the scores on a particular item and the total score on all other items, the better discriminator the item is. Kline (2000) suggests that a good item-total score correlation coefficient must be at least 0.3. The item-total correlation ensures that the test is homogeneous i.e. all items measure the same variable. (However validity studies are required to show what that variable is).

- If both the items' scores and the total score are continuous random variables, then the Product Moment Correlation Coefficient can be used instead of D.
- If the items' scores are dichotomous or can be dichotomized, then the Point Biserial or the Biserial Coefficients should be used respectively.
- If both the items' scores and total scores are dichotomized then we calculate the phi (ϕ) or the tetrachoric coefficients. (Howell, 1992: 265-283).

Howell (1992) explains in detail the methods for calculating these coefficients and the advantages and disadvantages of each.

One should be cautious when interpreting item analysis data because they cannot be used by themselves to judge the validity of a test and are influenced by factors like: the

number of items in the test, the nature and size of the group being tested, the instructional procedures employed by the teachers, chance error and the position of an item in the test.

Item analysis data provide a valuable service in selecting good test items. But they should be used as a 'flag' to identify items that may require more careful examination rather than as a shovel to bury suspect items.

(Mehrens and Lehmann, 1991: 168)

2.1.8 Item Response Theory (IRT)

Limitations of Classical Test Theory (CTT)

Hambleton, Swaminathan and Rogers (1991) identify the following limitations of CTT:

- Ability scores are item dependent (i.e. they depend on the item difficulty)
- The item statistics (difficulty, discrimination, reliability) are examinee dependent. Discrimination indices as well as reliability estimates tend to be higher in heterogeneous examinee groups than in homogeneous ones.
- No information is available about how examinees of specific abilities might perform on a certain test item
- Equal measurement error is assumed for all examinees (this measurement error is item dependent too)
- Classical item indices are not invariant across subpopulations (i.e. different subgroups of the sample of examinees give different item statistics).
- Reliability estimates assume parallel tests which in practice is difficult to satisfy.

According to Hambleton, Swaminathan and Rogers (1991) IRT provides alternative models, which have the following desirable features:

- Item characteristics that are not group dependent
- Scores describing examinees' abilities that are not test dependent
- A measure of precision for each ability score
- The probability that an examinee of any ability will answer items of any difficulty correctly.
- Do not require strictly parallel tests for assessing reliability.

The basic idea, around which IRT was developed, is that the probability of an answer given by a person to any item can be described as a function of the person's position on the latent trait (or ability measure) and one or more parameters.

Given the answers of n persons to k items, which are intended to measure the same latent trait, the person and item parameters can be estimated. Also the assumptions underlying the IRT model can be tested. This will help the researcher to:

- assess how good a measurement instrument the test is.
- predict the test's performance in future applications
- improve the quality of the test by indicating which items are inappropriate and should be changed or deleted.
- improve the quality of measurement by recognizing persons whose response pattern is unusual and their test scores may not be a valid measure of their position on the latent trait.

In IRT the probability of a correct response on an item is expressed as a function of the latent trait value θ and a number of item characteristics. Often used models are the 1-parameter, 2-parameter and 3-parameter logistic models (Baker, 1985; Hambleton et al., 1991; Hambleton, 1993; van der Linden and Hambleton, 1997). The 1-parameter model is often unhelpfully identified with the Rasch model because the mathematical function is the same. However, as it will be explained later, the major difference is in the philosophy of how and why the models were derived.

Two-Parameter Logistic Model

The two-parameter logistic model was proposed by Birnbaum. It is an item response model in which:

$$P_i(\theta) = \frac{\exp(Da_i(\theta - \beta_i))}{1 + \exp(Da_i(\theta - \beta_i))} \quad (i = 1, 2, \dots, n)$$

where $P_i(\theta)$ is the probability that a randomly selected examinee with ability θ will answer item i correctly, β_i is the difficulty index and represents the point on the ability scale at which the examinee has a 50% probability of answering item i correctly, a_i is the item discrimination and is proportional to the slope of $P_i(\theta)$ at the point where $\theta = \beta_i$. D is a scaling factor ($D = 1.7$) used to bring the interpretation of the parameters of the logistic model in line with those of the two-parameter normal ogive model which was the first model used before the logistic models..

The item characteristic curve is a monotonically increasing function specifying that as the level of the ability increases, the probability of a correct response to an item increases.

Three-Parameter Logistic Model

The three-parameter logistic model is obtained from the 2-parameter model by adding a third parameter c_i . It is an item response model in which:

$$P_i(\theta) = c_i + (1 - c_i) \frac{\exp(Da_i(\theta - \beta_i))}{1 + \exp(Da_i(\theta - \beta_i))} \quad (i = 1, 2, \dots, n)$$

where c_i is the lower asymptote of the item characteristic curve and represents the probability that examinees with low ability will answer the item correctly. It is considered like a guessing parameter, called by Hambleton et al. (1991, p.17) *pseudo-chance-level parameter* because typically it assumes values that are smaller than the values that would result if the examinees guessed randomly on the item.

Assumptions

The validity of the results of any statistical model is based on the specific assumptions about the data and the degree to which they are met.

The two main assumptions that should be met by the data are those of *unidimensionality* and *local independence*, which are described in more detail in the chapter on the assumptions of the Rasch model.

However, when using the 2-parameter logistic model, another assumption should be met too. That is, examinees with low abilities do not respond to an item correctly by guessing. This is inherent in the formula, because for all items with $\alpha_i > 0$ the probability of a correct response to the item decreases to zero as ability decreases.

2.2 The Rasch Models

2.2.1 The Dichotomous model (The Rasch Model)

One of the major problems in education and the social and behavioural sciences is that the performance of a person is not independent of the measuring instrument employed. This is inevitable because of the interaction between the person being measured and the instrument involved.

In the 1950s Danish Mathematician Georg Rasch saw that, although he could not determine exactly how a candidate would respond to an item, it should be possible to estimate the candidate's probability of success on that item. He also saw that, the probability for a right answer must only be governed by the candidate's ability (β) and the item's difficulty (δ).

The procedure, in which it is always the performance of a person relative to a particular item that is being considered in terms of probabilities, is called conjoint measurement. Thus, according to Masters and Keeves (1999), a person's ability is set at the same level as the item difficulty if that person has a specified probability (usually 0.5) of responding correctly to the item.

“The ability of the person and the difficulty of the item must be considered to be joined or conjoint in all analyses of responses and a principle of relativity with respect to the item must underlie the task of measurement. This principle overcomes the problems that were raised in earlier decades and that claimed that measurement was not possible in the social and behavioral sciences.”

(Keeves and Alagumalai, 1999, p. 25)

Rasch deduced the following formula for dichotomously scored performances:

$$\log \left(\frac{\text{Probability of success}}{\text{Probability of failure}} \right) = \text{Ability} - \text{Difficulty}$$

Then with simple mathematical steps he deduced the formula for a person n 's probability of scoring 1 rather than 0 on item i (P_{ni1}):

$$P_{ni1} = \frac{\exp(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)}$$

where β_n is the ability of person n and δ_i the difficulty of item i .

Property of invariance

Wright (1967) states that when this model governs measurement, one can free the item difficulty estimation from the abilities of persons in the calibration sample. At the same time ability estimation can be freed from the difficulties of the items used in the test.

Wright (1967) goes on to illustrate sample free measurement by means of two examples. He takes the worst-case scenario by choosing the two extreme groups from a sample of 976 students, the 'Dumb Group' (the 325 students with the lowest scores on the test) and the 'Smart Group' (the 303 students with the highest scores on the test). Item calibrations from the two groups give statistically equivalent item estimates, that is, the two estimates are close enough so that their differences are about what are expected from the uncertainty within the error of measurement.

He then obtained similar results for person measurement by dividing the 48 test items into two groups, the 24 easiest and the 24 hardest.

Wright and Masters (1982) argue that when a variable is used with different groups of persons or with the same persons on different occasions, it is essential that the variable maintains its identity from one measurement occasion to the other. 'Only if the item calibrations are invariant from group to group and from time to time can meaningful comparisons of persons be made' (p.114).

They then go on to describe ways of comparing item estimates from different calibrations giving in detail the method they prefer best, 'plotting estimates from different occasions' (p.115).

Given the item estimates from the two calibrating occasions, d_{Ai} (estimation of the difficulty of item i from the subset A) and d_{Bi} (estimation of the difficulty of item i

from the subset B) and their equivalent errors of calibration s_{Ai} and s_{Bi} then 95% confidence band can be constructed using

$$d_i \pm \sqrt{s_{Ai}^2 + s_{Bi}^2}, \text{ where } d_i = (d_{Ai} + d_{Bi})/2.$$

By plotting the points (d_{Ai}, d_{Bi}) together with the appropriate confidence band one can infer whether invariance holds. If substantially more than 5% of the points fall outside the confidence bands then that will provide evidence for a general lack of invariance.

The Rasch model is also a practical way to solve equating problems.

Data from different tests taken by different candidates can be combined and analyzed together, so long as there is some network of commonalities (candidates and/or items) linking the tests. This combined analysis provides a calibration, standard error and fit statistics for every item and a measure, standard error and fit statistic for every candidate involved in any of the testings. These item calibrations and candidate measures are completely equated because they are all expressed at once on one common linear scale. Once a bank of items has been calibrated, inclusion of items from the bank into each new test automatically equates that test to the common metric of the bank, and so to all other tests derived from the bank. (Wright, 1993, p.2)

Bond and Fox (2001) describe the basic principles of the Rasch Model and conclude the following:

- The Rasch model provides a mathematical framework against which test developers can compare their data.
- The model is based on the idea that useful measurement involves examination of only one human attribute at a time (unidimensionality) on a hierarchical line of inquiry.
- This line of inquiry is a theoretical idealization against which we can compare patterns of responses that do not coincide with this ideal. Person and item performance deviations from that line can be assessed, alerting the investigator to reconsider item wording and score interpretations from these data.

- Each item difficulty, and person ability is estimated on a logit scale, and each of these estimates has a degree of error associated with it, which decreases as information about difficulty and ability increases (i.e. items and persons are appropriately targeted).
- A logit value of 0 in item difficulty estimates is set arbitrarily as the mean of the difficulty estimates.
- Person ability is estimated in relation to the item difficulty estimates and
- Most Rasch software output include a form of item-person map in which person ability and item difficulty relations are easily seen. It is this item-person map that is very attractive to both experienced and new users.

The measurement unit in Rasch models is the logit, which simply means the log odds, that is, the natural logarithm of the probability of success divided by the probability of failure.

A person's ability in logits is their natural log odds for succeeding on items of the kind chosen to define the scale origin or "zero". Thus the person's

probability P for succeeding on an item with difficulty $\delta = 0$ is $\frac{e^{\beta}}{1 + e^{\beta}}$ from

which their success odds are $\frac{P}{1 - P} = e^{\beta}$, the natural log of which is β .

Similarly, an item's difficulty in logits is the natural log odds for failure on that item by persons with abilities at the scale origin. The probability P of these persons with abilities at $\beta = 0$ of succeeding on an item with difficulty

δ is $\frac{e^{-\delta}}{1 + e^{-\delta}}$ from which their odds for failure are $\frac{1 - P}{P} = e^{\delta}$, the natural log of which is δ .

(Wright, 1977, p.99)

As with all interval scales the origin of the scale is indeterminate. However, since it is the difference ($\beta - \delta$) which governs the probability of a right answer, we can add or subtract any constant to all abilities and difficulties without changing the bearing of their difference on the probability of success. Therefore, the origin is usually arbitrarily set to the average item difficulty for convenience.

Smith (2000) quotes the answer to a question raised in a discussion in the Rasch Sig Business meeting conducted at the 1999 American Educational Research Association. The question was about how one explains what a logit is to non-Rasch practitioners. The answer was “Who cares what a logit is as long as you find it useful”.

Smith’s hope was not so much to help researchers understand the technical definitions of the logit metric but to help them realize its usefulness.

2.2.2 Rasch model derived from objectivity

(Wright, 1988; Maters, 2001; Wright and Linacre, 1987)

Thurstone (1928) states (as quoted in Wright, 1988, para. 3):

The scale must transcend the group measured... A measuring instrument must not be seriously affected in its measuring function by the object of measurement. To the extent that its measuring function is so affected, the validity of the instrument is impaired or limited. If a yardstick measured differently because of the fact that it was a rug, a picture, or a piece of paper that was being measured, then, to that extent the trustworthiness of the yardstick as a measuring device would be impaired. Within the range of objects for which the measuring instrument is intended, its function must be independent of the object of measurement.

Thurstone is setting the grounds for objectivity. Objectivity is the requirement that the measures produced by a measurement model must be sample free for the items and test free for the people.

Essential to the concept of measurement is that of comparison. A model is required for comparing and hence estimating the position of two persons n and m on the ability scale independently of the items used to provide evidence of their relative standings on the scale.

For a test consisting of homogeneous items we do expect that the ratio of the count of right answers to that of wrong answers will remain approximately constant no matter what the length of the test was.

Consequently, a ratio is the type of comparison for which we desire to construct measures.

Hypothetically, if an item is repeatedly administered numerous times to the same two hypothetical persons n and m , who answer each question independently, then the following table would result:

	Person n	
Person m	$n_r m_r$	$n_w m_r$
	$n_r m_w$	$n_w m_w$

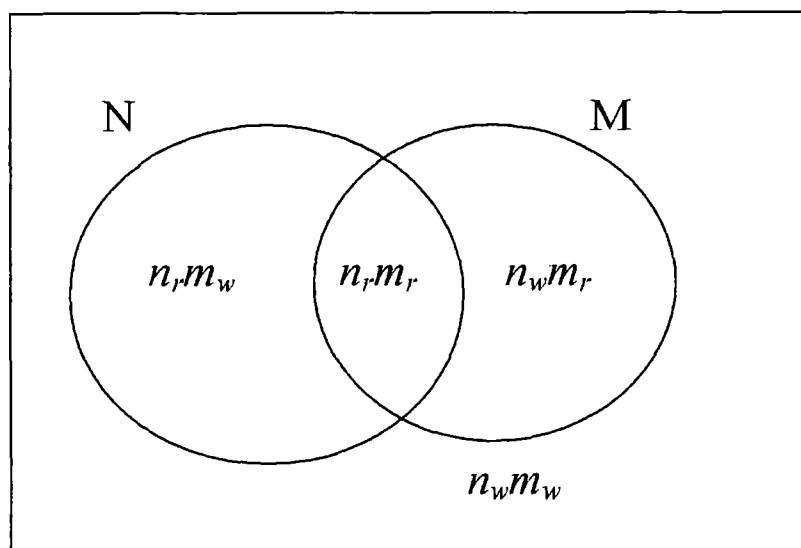
Where $n_r m_r$ is the count of times when both persons answer the item correctly

$n_w m_r$ is the count of times when n answers the item incorrectly and m correctly

$n_r m_w$ is the count of times when n answers the item correctly and m incorrectly

$n_w m_w$ is the count of times when both persons answer the item incorrectly.

The same information can also be displayed in a Venn diagram as shown below.



where N is the event 'person n answers the item correctly' and M is the event 'person m answers the item correctly'.

The only two counts that contain information useful for comparisons of the performances of the two persons are $n_w m_r$ and $n_r m_w$.

The ratio $\frac{n_r m_w}{n_w m_r}$ is a comparison of the frequencies of success of the two persons on the item in question. This is the ratio we want.

If we divide both numerator and denominator of this ratio by $n(\Omega)$, the number of times the item is administered to persons n and m we get:

$$\frac{n_r m_w}{n_w m_r} = \frac{\frac{n_r m_w}{n(\Omega)}}{\frac{n_w m_r}{n(\Omega)}} = \frac{P(N \cap M')}{P(N' \cap M)}$$

Hence, since the events N and M are independent
$$\frac{n_r m_w}{n_w m_r} = \frac{P(N) \cdot P(M')}{P(N') \cdot P(M)}$$

and writing this in a slightly different notation:

$$\frac{n_r m_w}{n_w m_r} = \frac{p_{ni} \cdot (1 - p_{mi})}{(1 - p_{ni}) \cdot p_{mi}} \quad (1)$$

Where p_{ni} is the probability of success of person n on item i and

$1 - p_{ni}$ is the probability of failure of person n on item i.

Using objectivity, the comparisons of the performance of the two persons must be independent of which items are used. Therefore, the ratio of the comparison must be the same for any two items i and j, giving:

$$\frac{p_{ni} \cdot (1 - p_{mi})}{(1 - p_{ni}) \cdot p_{mj}} = \frac{p_{nj} \cdot (1 - p_{mj})}{(1 - p_{nj}) \cdot p_{mi}} \quad (2)$$

Multiplying both sides by $\frac{p_{mi}}{1 - p_{mi}}$ gives

$$\frac{p_{ni}}{1 - p_{ni}} = \frac{p_{nj} \cdot (1 - p_{mj})}{(1 - p_{nj}) \cdot p_{mj}} \cdot \frac{p_{mi}}{1 - p_{mi}} \quad (3)$$

For simplicity let $j = 0$ and $m = 0$ be the origins for the item scale and the person scale respectively.

This makes the measure of person n its difference from the 'standard' person $m = 0$ and the calibration of item i its difference from the 'standard' item $j = 0$.

Then equation (3) becomes:

$$\begin{aligned} \frac{p_{ni}}{1 - p_{ni}} &= \frac{p_{n0} \cdot (1 - p_{00})}{(1 - p_{n0}) \cdot p_{00}} \cdot \frac{p_{0i}}{1 - p_{0i}} \Rightarrow \\ \frac{p_{ni}}{1 - p_{ni}} &= \frac{p_{n0}}{1 - p_{n0}} \cdot \frac{p_{0i}}{1 - p_{0i}} \cdot \frac{1 - p_{00}}{p_{00}} \end{aligned} \quad (4)$$

where

$\frac{p_{n0}}{1 - p_{n0}}$ is the ratio of the probability of success of person n on the 'standard' item 0 to the probability of failure of person n on the 'standard' item 0.

$\frac{p_{0i}}{1 - p_{0i}}$ is the ratio of the probability of success of the 'standard' person 0 on item i to the probability of failure of the standard person 0 on item i .

$\frac{1 - p_{00}}{p_{00}}$ is the ratio of the probability of failure of the 'standard' person 0 on the

'standard' item 0 to the probability of success of 'standard' person 0 on the 'standard' item 0.

If we bring the frame of reference for persons and items into conjunction by choosing

the reference (standard) item and person such that $p_{00} = 0.5$, then $\frac{1 - p_{00}}{p_{00}} = 1$.

Therefore equation (4) becomes:

$$\frac{p_{ni}}{1 - p_{ni}} = \frac{p_{n0}}{1 - p_{n0}} \cdot \frac{p_{0i}}{1 - p_{0i}} \quad (5)$$

The measurement scale now defined by $\frac{p_{ni}}{1 - p_{ni}}$ has the properties of a ratio scale with:

$$0 < \frac{p_{n0}}{1 - p_{n0}} < \infty \quad \text{depending only on person } n \text{ and}$$

$$0 < \frac{p_{0i}}{1 - p_{0i}} < \infty \quad \text{depending only on item } i.$$

This ratio scale can now be transformed into linear form by:

$$\ln\left(\frac{p_{ni}}{1 - p_{ni}}\right) = \ln\left(\frac{p_{n0}}{1 - p_{n0}}\right) + \ln\left(\frac{p_{0i}}{1 - p_{0i}}\right) \quad (6)$$

And if we let $\ln\left(\frac{p_{n0}}{1 - p_{n0}}\right) = B_n$ & $\ln\left(\frac{p_{0i}}{1 - p_{0i}}\right) = -D_i$ then equation (6) becomes

$$\ln\left(\frac{p_{ni}}{1-p_{ni}}\right) = B_n - D_i \Rightarrow \frac{p_{ni}}{1-p_{ni}} = \exp(B_n - D_i) \Rightarrow$$

$$p_{ni} = \exp(B_n - D_i) - p_{ni} \cdot \exp(B_n - D_i) \Rightarrow$$

$$p_{ni}[1 + \exp(B_n - D_i)] = \exp(B_n - D_i) \Rightarrow$$

$$p_{ni} = \frac{\exp(B_n - D_i)}{1 + \exp(B_n - D_i)}$$

where the item calibration D_i is dependent only on the attributes of item i and B_n is the person measure depending only on the attributes of person n .

And this is the Rasch Model, the only IRT model derived from objectivity.

2.2.3 Assumptions – Model fit

Statistical models usually base the validity of their results on the specific assumptions about the data. Violations of these assumptions can cause failure of the model invalidating the results of the analysis.

Unidimensionality

An assumption common to the most widely used Item Response Theory (IRT) models is that the items that make up the test measure only one ability. This is called the assumption of *unidimensionality*.

According to Smith Jr. (2004b, pp 575-576), Stout (1987) states that there are at least three reasons why it is important that responses to an assessment represent a

unidimensional construct. First, any measure of the level of a construct should not be influenced by varying levels of one or more other abilities. Second, an assessment to be used in identifying differences or ordering persons on some attribute must measure a unidimensional construct. This is a requirement for two persons with the same score to be considered similar. Third, unidimensionality must hold before the total score is calculated or the ability estimated, as violations of this requirement may bias item and person estimates.

Unidimensionality is an essence of measurement. In fact one of the reasons that make the Rasch model so important as the method for constructing measures is its deduction from the requirement of unidimensionality.

Wright and Linacre (1989) admit that in practice no test can ever be perfectly one-dimensional. Nevertheless the ideal of unidimensional measures must be approximated if generalizable results are to be obtained.

Hambleton et al. (1991) also state:

What is required for the unidimensionality assumption to be met adequately by a set of test data is the presence of a dominant component or factor that influences test performance. This dominant component or factor is referred to as the ability measured by the test. (pp. 9-10)

Smith Jr. (2004b) gives a similar description of unidimensionality, in the context of the Rasch model and the trait estimates:

Essential unidimensionality is based on the premise that a dominant dimension exists with the possible presence of several minor dimensions and that the dominant dimension is so strong that the trait estimates are not affected by the presence of the smaller dimensions. (p. 577)

Often constructs of interest in the social sciences are complex and are represented by a set of correlated factors.

According to Athanasou and Lamprianou (2002), Bejar (1983) suggested that unidimensionality did not necessarily mean that the performance on the questions was due to a single cognitive process. Instead he proposed that a variety of cognitive processes could be involved as long as they functioned in unity. Therefore "it is

possible to fit the Rasch model on the results of a test that actually measures a few highly related abilities.”(Athanasou and Lamprianou, 2002, p.223).

Also, Masters and Keeves (1999), in describing the strengths of the Rasch models state that unidimensionality is “no longer a restriction, provided that a limited number of dimensions have been hypothesized, and the items and persons are constrained to these dimensions.” (p. 13)

In suggesting how unidimensionality can be achieved Wright and Linacre (1989) suggest that the pursuit of unidimensionality is undertaken at two levels. First, the test items, tasks, observation techniques and other aspects of the testing situation should be organized to realize, as perfectly as possible, the variable which the test is intended to measure. Second, the test analyst should collect a relevant sample of these carefully designed observations and evaluate the practical realization of that intention.

Assistance in examining the unidimensionality of a set of test items is provided by the fit statistics, which report the degree to which the observations meet this vital specification for measurement. Under Rasch analysis, if all items cohere to a single scale unidimensionality may be asserted. Misfitting items can be redesigned or replaced.

Every time we use our measuring agents, questions or items to collect new information from new persons in order to estimate new measures we must verify in those data that unidimensionality requirements of our measuring system have once again been sufficiently well approximated to maintain the quantitative utility of the measures produced.

(Wright and Linacre, 1989, p.7)

Local independence

Another main assumption of the Rasch model and other IRT models is the assumption of *local independence*.

Local independence means, “when the abilities influencing test performance are held constant, examinees’ responses to any pair of items are statistically independent” (Hambleton et al., 1991, p.10).

Simply put, it means that the response of a person to a question should not affect responses to other questions. For example, previous questions should not give hints or insights for the solution of the next questions.

Other than the unidimensionality and local independence, the Rasch model requires **three more assumptions.**

First, the test is expected to be a power test, that is, the students should have enough time to attempt all the questions in the test. This assumption is a safeguard to unidimensionality because if the test is timed, then the speed of the examinee in grasping and handling tasks enters into the picture and the unidimensional structure of the tasks is distorted.

Second, minimal guessing is one factor that should always be checked before the use of the Rasch model. If there is a lot of successful guessing then items would not fit the model. Guessing is however usually only a problem with multiple choice or matching questions.

Third the Rasch model demands that the questions discriminate between the more and the less able students in a similar way. Linacre (1996) states that control misfit statistics flag items that fail to meet this measurement specification.

Because of all these assumptions, it is harder to create a test constrained by the requirements of Rasch measurement than it is to construct a classical test.

2.2.4 Comparing the 2-P and 3-P models with the Rasch model

Wright (1983) argues that fundamental measurement in the social sciences is obtainable only through the Rasch model and, in comparing the Rasch model with the 2-P and 3-P models, states:

If measurement is our aim, nothing can be gained by chasing extra item parameters like c and a . We must seek, instead, for items which can be managed by an observation process in which any potentially misleading disturbances which might be blamed on variation in possible c 's and a 's can be kept slight enough not to interfere with the maintenance of a scale stability sufficient for the measuring job at hand. ... Only the Rasch process can maintain units that support addition and so produce results that qualify as fundamental measurement. (Wright, 1983. p. 7)

Furthermore, the Rasch model is the only one which uses the raw score as the sufficient statistic for estimating item difficulty or person ability. That is, the sufficient statistic for estimating person ability is the sum or count of the correct responses for a person over all items. Similarly, the sufficient statistic for estimating item difficulty is the sum or count of the correct responses for an item over all persons.

In the other two models the sufficient statistic for ability estimation includes other parameters that must be estimated simultaneously.

Wright (1995) compares the Rasch model with the 3-parameter model using the 1992 National Adult Literacy Survey (NALS) with 24944 adult participants and 173 literacy items. He shows that the 3-P discrimination is highly and negatively correlated ($r = -0.82$) with the infit mean statistics (when both are log-scaled) and argues that to find the 3-P discriminations in a Rasch analysis one only needs to look at the infit mean square statistics.

He then shows that by plotting the 3-P lower asymptotes (guessing parameters) against the outfit statistic almost no guessing has occurred (which would have been detected by outfit), except from 2 out of the 13 multiple choice items. He concludes that:

The bulky and complex NALS data, containing a wide variety of dichotomous item types and administered to a large and diverse sample of

respondents, is just the data expected to manifest all the features that would make the superiority of the 3PL clear. This parallel NALS analysis shows, however, that 3PL has no benefits over Rasch and some detriments. 3PL ability estimates and item difficulties are equivalent to Rasch measures. 3PL item discrimination provides the same information as the Rasch infit statistic, but parameterising item discrimination complicates estimation. It also inhibits interpretation and use of item difficulties by obscuring the item hierarchy and hence the construct definition. (Wright, 1995, p.408)

His final remarks are on guessing and he claims that including a lower asymptote can be harmful. In most cases, there is no lucky guessing, so adding this parameter penalizes all respondents, particularly the lower performers who really knew the answer. He suggests that in the few cases where guessing is actually thought to have occurred one can remove the easily detectable assumed guesses from the data set, treating those few items as not administered to those few people. This way only those who have guessed are penalized, and only by the very small amount by which their lucky guessing boosted their performance.

In comparing the 2-parameter and 3-parameter models with the Rasch model it is important to distinguish between measurement and modeling. If the purpose is to construct a good measure then the items and the test should be constrained to the principles of measurement. If on the other hand the purpose is to model some test data then the model which fits the data best should be chosen. Rasch corresponds to the principles of measurement whereas other IRT models correspond to modelling. In the latter case Fischer and Molenaar (1995) state that:

They (the 2-p and 3-p models) make less stringent assumptions (than the Rasch model), and are therefore easier to use as a model for an existing test. On the other hand, they typically pose more problems during parameter estimation, fit assessment and interpretation of results. Whenever possible, it is thus recommended to find a set of items that satisfies the Rasch model rather than find an IRT model that fits an existing item set.

(Fischer and Molenaar, 1995, p.5)

Here they are taking a modelling perspective and conclude that the Rasch model is best.

Linacre (1996) adds to the above that allowing or parameterising discrimination or guessing, which are sample dependent indices, limits the meaning of the measures to just that subset of items and persons producing these particular data. This prevents any general inferences over all possible items probing that construct among all possible relevant persons.

Another important difference is the sample sizes required for the calibrations. The use of the 2-P or 3-P models requires larger samples of persons for calibrations. Thissen and Wainer (1982) have worked out a complicated formula for obtaining the standard errors of the parameters estimated, as a function of sample size and the parameters, for any logistic item response model when the maximum likelihood method of estimation is used. According to their formula, the 1-P and 2-P models give approximately the same standard errors for item difficulties very close to 0 logits, (using a slope of 1.5, i.e. discrimination of 1.5) when a sample of 2500 is used.

In a further example to show how their formula can be used to find the sample size required to give an accuracy of one decimal place (i.e. standard error of location of 0.05) they used a slope parameter of 1 – 1.5 (considered good test items) and items with locations from – 2 to 2 logits. In the worst case situation (item locations close to – 2) for the 1-P model a sample of size 2500 was needed whereas for the 2-P and 3-P models the equivalent sample sizes were 7500 and 67000. In concluding they state:

... try to fit the simplest models first, and only when they are found to be inadequate move on (with trepidation) to the more complex ones. If the more complex models are required it would seem that a method other than unrestricted maximum likelihood ought to be used.

(Thissen and Wainer, 1982, p. 407)

Masters and Keeves (1999) note that simplicity and generality are the benefits in using the Rasch model and identify a possible disadvantage of the Rasch models, that of the exclusion from calibration of non-fitting persons. They conclude however that:

“Estimates of person performance may nevertheless be made for those persons excluded, and advantages are gained through improved measurement.” (p. 13)

In conclusion, the Rasch model is based on a different philosophy from the other approaches. This philosophy dictates the structure of the data including the fact that unidimensionality is a must for the measurement process. The other models are driven by a desire to model all of the characteristics observed in the data, regardless of whether they have any contribution to the measurement process.

2.2.5 Discrimination again: Is higher discrimination always better?

In Classical Test Theory (CTT) high discrimination is considered a desirable characteristic of an item and a strong indication of its quality. In fact, the higher the discrimination the better the item is. The reason for this special importance placed on highly discriminating items stems from the use of psychological and educational tests for purposes of separating individuals by ability or by their position on the latent trait.

Masters (1988) argues that item response models that incorporate a discrimination parameter (such as the 2-P and 3-P models) also treat highly discriminating items as the best items on the test.

In the estimation of abilities in the 2-P parameter model, for example, the sufficient statistic is:

$$r_n = \sum_{i=1}^L x_{ni} \alpha_i$$

where r_n is the estimated ability of person n , α_i is the estimated discrimination of item i ($i = 1, 2, 3, \dots, L$) and x_{ni} takes the values 0 or 1 depending on whether the response to item i was wrong or correct respectively.

This leads to success on a highly discriminating item always being worth more than success on a less discriminating item (i.e. the higher an item's discrimination the higher its influence on estimates of ability is)

Under these approaches to test construction and revision, unusually discriminating items are sought after and attempts are made to write more items like the highly discriminating items already developed. Provided that they display adequate face, or content, validity and are of appropriate difficulty, these are the last items a test constructor is likely to be concerned about when reviewing a test, and the last items likely to be inspected for possible flaws. (Masters, 1988, p. 16)

The Rasch perspective

Items satisfying the requirements of the Rasch model must be of about equal discrimination. According to Masters (1988) the items that CTT identify as best and other IRT models give greatest weight in the measurement process, the Rasch model identifies as problematic. 'This feature of the Rasch model is a significant departure from established practice and challenges a fundamental tenet of popular item analysis' (Masters, 1988, p.16)

He then argues that the very items that test constructors might otherwise have believed were the best in their test are identified by the Rasch model as suspect and in need of investigation and describes the following cases where high discrimination is problematic.

Different item performance

A form of differential item performance can be of a special concern in some settings if it results from differences in opportunities to learn the content of particular test items in different instructional programmes.

For example this situation may arise if students were divided into two different instructional levels based on their abilities, say level A (lower demanding) and level B (higher demanding).

At the end all students may take the same test or some common items, for test equating purposes. If the content of a specific item had been taught thoroughly to the students of the higher ability level (B) but not taught or treated superficially to the students of the lower ability group (A), then this would result in that item being highly discriminating, perhaps the most discriminating item in the test. The reason for this is that it provides

level B students with a special advantage and the item reflects differences in opportunities to learn, which in this case happens to be highly correlated with ability.

Opportunity to answer

In a speeded test traditionally items answered incorrectly and items not attempted are treated in the same way, both scored as wrong.

In general, examinees that reach the last items and have time to attempt them are likely to be the more able persons in the group. This means that the examinees of low ability may suffer a special disadvantage that would perhaps have not suffered if the same items were presented in isolation or at the beginning of the test.

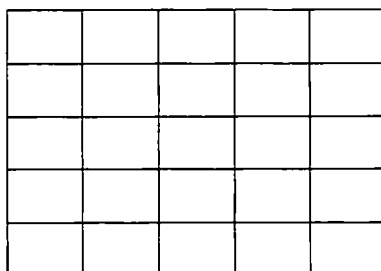
The effect will be to make the item more discriminating.

Test wiseness

The occasional item that is sensitive to differences in test wiseness is likely to favour students who are already at an advantage because of their higher ability, and may operate against the lower ability students making the item unusually discriminating.

As an example Masters (1988, pp 27 – 28) gives the following maths item which was administered to a group of 14-year-old students in San Antonio Texas.

How many squares are there in this 5" by 5" grid?



(Right answer = $1^2 + 2^2 + 3^2 + 4^2 + 5^2 = 55$)

The existence of an obvious but incorrect answer to the item (25 squares) appears to have prevented less able or more naïve students from engaging with the intended task and thus setting them at a special disadvantage.

Because this second dimension (test wiseness) will in general be positively correlated with ability such items will be more discriminating.

Another case which could give rise to problematic high discrimination is what the researcher calls **special knowledge in favour of high ability persons**.

The following example is given by Masters (1993):

In a second-language comprehension test in German, for the Dutch National Institute of Educational Measurement, each listening item is given to Dutch students and German native speakers of the same age and test results are Rasch analysed.

There was on one occasion, in 1987, an unusually discriminating item, an excerpt from German radio. Native speakers (overall high performers) did unusually well relative to the Dutch students (overall lower performers). An inspection of the item showed that it was based on a conversation about German politics. The native speaking (German) students had an advantage on this item because of their ordinary knowledge of German politics.

This is another example where an item is highly discriminating because of its sensitivity to a second irrelevant dimension that is highly correlated with the variable of interest. 'The contaminating influence of a second dimension often manifests itself in unusual item discrimination'. (Masters, 1993, p. 289).

In concluding, Masters (1988) states:

Secondary influences that operate to give persons of high ability a special advantage on an item may be subtle. ... The first step in their identification is the recognition that unusual item discrimination can be an indication that an item is giving some individuals an unintended advantage. The responsibility then lies with the test developer to investigate each unusually discriminating item to determine whether or not it is introducing and giving special weight to differences on a second undesired dimension. (pp. 28 – 29).

The Rasch model identifies items with unusually high discriminations and cautions test developers to the possible existence of the above mentioned problem.

Linacre (personal communication, March 27, 2008) also argues that a highly discriminating item could be acting as a summary of other items. It is not acting as an independent item and although in most cases this will not substantially matter, it may reduce the efficiency of the test as a measuring instrument.

2.2.6 Rasch polytomous models

This section introduces the Rating scale and Partial Credit models and compares them by explaining the similarities and differences of the two models, as well as their applicability.

The Partial Credit Model

The original model developed by Rasch was for the analysis of responses, which are scored dichotomously. However in educational assessment the multistep problems are very common particularly in subjects like Mathematics and Science. These items are designed to assess students' abilities to identify an appropriate solution strategy and to pursue this strategy to a successful conclusion. In these items it is common to award partial credit, for partial success, in the hope that this will lead to more precise estimates of persons' abilities.

The model

Masters (1982) in his introduction of the partial credit model states that 'when performances on an item are recorded in the $m + 1$ ordered levels 0, 1, 2, ..., m , it is convenient to think in terms of the m steps which have to be taken to complete the test.'

He then introduces the Partial Credit model (PCM), which is given by:

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

where for notational convenience $\sum_{j=0}^0 (\beta_n - \delta_{ij}) = 0$, and

Π_{xni} is the probability of a person n scoring x on item i ,

β_n is the person's position on the variable

δ_{ij} are the difficulties of the m_i 'steps' in item i .

Bode (2004) describes three situations in which the PCM can be used.

First, when instruments contain items with varying degrees of correctness for responses that can be ordered from least correct to most correct, like a multiple-choice test used to measure reading comprehension in which some responses might be more correct than others.

Second, when instruments contain items that can be broken into component tasks, the first of which must be completed before the next is attempted, and each of which can be scored as correct or incorrect like scoring constructed responses measuring complex mathematical problems.

Third, when instruments contain items where increments in the quality of a performance are rated, like a student history portfolio that is rated on a number of criteria.

The Rating Scale Model

The Rating Scale Model (RSM) belongs to the family of Rasch models and is a special case of the polytomous model.

"The main assumption for the RSM, apart from being a polytomous Rasch model, is that scoring of the response categories must be equidistant, i.e. their values must increase by a constant" (Andersen, 1997, p. 67).

The model

Masters and Wright (1982) also present this model. The probability of a person n responding in category x to item i , is given by:

$$P_{xni} = \frac{\exp \sum_{j=0}^x [\beta_n - (\delta_i + \tau_j)]}{\sum_{k=0}^m \exp \sum_{j=0}^k [\beta_n - (\delta_i + \tau_j)]} \quad x = 0, 1, \dots, m$$

where $\tau_0 = 0$ so that $\exp \sum_{j=0}^0 [\beta_n - (\delta_i + \tau_j)] = 1$, and

β_n is the person's position on the variable

δ_i is the scale value (difficulty to endorse) estimated for each item i and

$\tau_1, \tau_2, \dots, \tau_m$ are the m response thresholds estimated for the $m + 1$ rating categories.

The RSM requires that all the items in a test have the same number of steps, as we would expect for example from Likert scales in attitude instruments. This model is widely used for the analysis of Likert scales, even though the original intention of Andrich, according to Bond and Fox (2001), was to use it in the evaluation of written essays.

PCM Vs RSM

The PCM and the RSM are very similar in that they both share the assumptions of unidimensionality, local independence and minimal guessing and the same statistics, that is, ability and difficulty estimates, error of estimates and mean square fit statistics to evaluate the quality of measurement.

Just as the PCM is an extension of the dichotomous model, the RSM is a simplification of the PCM. In the PCM, the transition from one category to the next can have a different meaning from one item to another. In contrast, the RSM forces a single scale structure on the responses across all items.

In simpler words, in the PCM each item may have a different number of steps or categories and each step can have a different difficulty estimate from item to item, whereas in the RSM the same category has exactly the same meaning across the items.

In terms of the applicability of the models, the PCM is primarily used for achievement tests but the RSM with questionnaires and other rating scales.

Applications of the models

Rasch measurement has been applied in very diverse situations and some examples are outlined below:

Prieto, Roset and Badia (2001) have used the Rasch dichotomous model to assess the metric properties of the Spanish version of the assessment of Growth hormone deficiency in adults and to confirm its unidimensionality and construct validity.

Bond and Fox (2001) describe how data from Piagetian interviews have been analysed using the Rasch approach to give fresh insights.

Lee and Fischer (2005) evaluated the psychometric properties of the diabetes self-care scale (DSCS). Although the construct validity of the DSCS was supported by the analyses, Lee and Fischer made a few recommendations for improving the scale. Two of those recommendations were: (a) to add 10 more items which would be more difficult to endorse in order to differentiate better between people with extremely high level of self care and (b) to modify the categories from a 6-point rating scale to a possibly 3- or 4-point rating scale followed by further confirmatory analysis.

Massof and Fletcher (2001) have used the Rasch model to evaluate the validity of and to improve the visual functioning questionnaire which is designed to assess health-related quality of life of patients with visual impairment. Their analyses showed that the 17 items that require difficulty ratings produced a valid interval scale for low vision patients whereas the 10 items that require frequency or level of agreement ratings do not work together to produce a valid interval scale.

Chen, Bezruczko and Ryan-Henry (2006), driven by the need of health and social agencies to have systematic means of describing mothers' effectiveness in caregiving for their adult children with intellectual disabilities, have found through Rasch analyses, 61 items defining the empirical construct 'Functional Caregiving'. Those 61 items also defined 3 caregiving levels: advocacy, personal caregiving and community.

Myford and Wolfe (2002) examined a procedure for identifying and resolving discrepancies in ratings whereas, Lamprianou (2006) investigated the stability of two marker characteristics across tests: (a) severity and (b) consistency of marking. In both of the above-mentioned studies the many-facets Rasch model was used.

The above selection of applications of the Rasch models show the diversity of situations in which the Rasch models can be used productively over and above the usual assessments of ability in educational tests, the position on the latent trait in psychological tests or the identification of aberrant responses in tests or psychometric scales.

2.2.7 Criticisms of the Rasch models

2.2.7 (i) *Rasch's different approach to the data-model relationship*

Although the exponential models were known by the time Rasch worked with them he did not use them in the traditional way. Instead of investigating whether the models could fit a given set of data, he had the insight to make a case for them independently of any data and to argue for a different data-model relationship from the traditional.

Traditionally, the choice of one model over another is based on whether it accounts better for the data. In other words the choice of accepting or rejecting concerns the models and is based on the given data.

But as Andrich (2004) notes, the reason that Rasch's model turns the traditional data-model relationship upside down is that the model does not describe any data. "The model renders in mathematical, and most importantly from a practical and applied prospective, testable form, the requirements of measurement" (p. 172). Andrich is referring to the requirements of invariant comparisons, on which Rasch based his mathematical derivation of the model and quotes Rasch (1961) summarizing those requirements:

The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which other stimuli within the considered class were or might have been compared.

Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for comparison; and it should also be independent of which other individuals were also compared on the same or some other occasion.

(Andrich, 2004, p. 173)

Andrich (2004) argues that it is this fundamentally different approach to the data-model relationship that is resisted and from which the many criticism of the Rasch model have originated. He equates the Rasch approach to a paradigm shift of the type identified by Kuhn (1970) and draws parallels with other paradigm shifts and the criticisms that they drew from “experts” at the time only to become orthodox later.

2.2.7 (ii) *The criticisms*

One of the people who strongly opposed the use of the Rasch model in the UK, in the late 70s was Goldstein. In an article, in 1979 he outlined several criticisms of the Rasch model. Dickson and Kohler (1996) also expressed several criticisms in commenting on their analyses of the Functional Independence Measure (FIM) ratings. (FIM records the severity of disability of rehabilitation patients).

Between them, Goldstein (1979) and Dickson and Kohler (1996) cover the majority of the criticisms against the Rasch model, and responses to their criticisms are given below.

Others have criticized the Rasch model also, like Divgi (1986, 1989) who claimed that the model was not appropriate for multiple-choice items and like Whitely and Dawis (1974) and Whitely (1977), who criticized technical aspects of the model like estimation procedures and sample sizes.

Criticism 1: Unidimensionality

Goldstein’s (1979) first criticism, and probably the most frequently occurring one, refers to the assumption of unidimensionality and more precisely to the fact that in order to fit the Rasch model the items must “relate only to one underlying dimension of ability” (p.214). He differentiates the Rasch model from factor analysis (as methods for

detecting the dimensionality of data) in that in factor analysis “the dimensionality or number of factors is studied in the analysis itself” (p.214), implying the superiority of factor analysis. Dickson and Kohler (1996) also refer to the requirement of a one-dimensional latent space in their criticisms on the Rasch model.

Response to criticism 1

The measurement of any object in the physical sciences describes only one attribute of the object being measured. ‘This is a universal characteristic of all measurement’ (Thurnstone, 1931, p.257).

The importance of unidimensionality of a test is outlined by Stout (1987). He points out that it is important for a test that purports to measure the level of a certain ability not to be significantly contaminated by varying levels of other abilities displayed by the examinees taking the test. It is important that a test designed to be used in the measurement of individual differences must in fact measure a unified trait. Otherwise, it will be impossible to make valid inferences from the test results or to identify the individual differences.

Since Goldstein’s article, many psychometricians (see for example Hambleton et al., 1991; Masters & Keeves, 1999; Smith Jr., 2004b; Wright and Linacre, 1989) made it clear that unidimensionality does not implicitly mean only one factor or dimension but instead the presence of a dominant dimension with the possible presence of minor dimensions which do not affect the dominant one.

Hambleton (1993) clarifies that “the unidimensionality assumption cannot be strictly met because there are always other cognitive, personality and test-taking factors that affect test performance, at least to some extent” (p. 150). Possible factors include test motivation, test anxiety, speed of performance, test sophistication, reading proficiency and other cognitive skills. Hambleton (1993) concludes:

What is required for the assumption of unidimensionality to be met to a satisfactory extent by a set of test data is a *dominant* component or factor. ... This ability is broadly defined to reflect whatever the test measures: a cognitive ability, a measure of achievement, a basic competency or skill or a personality variable. What the ability is must be established in the

same way that the construct measured by any test is identified: through a construct validation investigation (p. 150)

According to Linacre (1998a), the presence of more than one dimension in the data does not necessarily imply substantive multidimensionality. Extra dimensions may reflect different person response styles or different item content area. For example, items on subtraction may define a different dimension than items on addition in a simple mathematics test for young children. Multidimensionality can also be an artifact of test construction. For example, including the identical item several times in a test produces a subset of highly intercorrelated items which may define an extra dimension. On the other hand, the use of different response mechanisms across items (multiple-choice, constructed-response, rating scales) introduces unmodeled variation which can be attributed to a dimension of 'item type'.

Multidimensionality only becomes a real concern when there are response patterns in the data indicating that the data represent two or more dimensions so disparate that it is no longer clear what latent dimension the Rasch dimension operationalizes. (Linacre, 1998a, pp 5-6)

As far as factor analysis is concerned, Linacre (1998a) showed that Rasch analysis followed by principal components analysis of standardized residuals was always more effective at both constructing measures and identifying multidimensionality than direct factor analysis of the original response-level data.

Principal components analysis of the standardized residuals is based on the specification of 'local independence', which is an assumption of the Rasch model. This asserts that, after the contribution of the measures to the data has been removed, what is left is random, normally distributed noise. Therefore the standardized residuals are modeled to have unit normal distributions which are independent and so uncorrelated. This is testable. If the resulting common factors explain nothing more than random noise across items, then the data conform to the Rasch model. The existence of substantive common factors, however, would indicate departure from unidimensionality.

“The aim of factor analysis of Rasch residuals is thus to attempt to extract the common factor that explains the most residual variance under the hypothesis that there is such a factor. If this factor is discovered to merely explain random noise, then there is no meaningful structure in the residuals.”

(Linacre 1998b, p. 636)

Criticism 2: The use of probabilities

Dickson and Kohler (1996) argue that any system of measurement based on probabilities must necessarily be imprecise.

Response to criticism 2

All measurement is made with error and an explicit acknowledgement that this is so can allow the researcher to express test success in probability terms. Even a ruler measurement is the most likely length of the object given the observation. The Rasch model does not introduce probabilities or imprecision into the data, on the contrary, it capitalizes on their presence in the data to construct a measurement system.

Criticism 3: The absence of distributional descriptions

Dickson and Kohler (1996) criticize also the fact that no description of the sample distribution exists in Rasch analysis.

Response to criticism 3

The Rasch model does not need to assume anything about the distribution of the sample. Parallels can be drawn with measures of weight and height and this is one of the strengths of Rasch measurement. It can reveal the underlying distribution. It is not dependent on assumptions about hypothesised distributions.

Criticism 4: Constancy of item difficulties

Goldstein (1979) refers to the fact that the relative difficulty of the items in a test is the same for all individuals. He states: “Hence, even if we were satisfied that a test tapped only one dimension of ability, in order to use the Rasch model we would also require that, despite different experiences, learning sequences etc., the difficulty order of items was the same for every individual” (p.214), implying that because of different

experiences, learning sequences etc. the difficulty order could not be the same for everyone.

Dickson and Kohler (1996) also criticise the assumption that item parameters are the same across all samples.

Response to criticism 4

Both Goldstein and Dickson and Kohler are referring to the property of invariance. This basic principle of order (or invariance) is not only an assumption of the Rasch model, but also the fundamental requirement for measurement.

Rasch, was not the first to require the same kind of invariance in social measurement. L. L. Thurnstone and L. Guttman, two of the most significant people in this field, both articulated this requirement. However, for Thurnstone this was only a property of the data, and although Guttman articulated a response structure to which data must conform, it was deterministic and most significantly it was not expressed in a mathematical form.

In a distinctive contrast with Thurnstone and Guttman, and reflecting Rasch's training as a mathematician and his instinct for mathematical rigour, Rasch built the properties of invariance into a class of mathematical models to which we now attach his name. This leads to another reason that the Rasch models can be subtle. Because the property of invariance is built into a mathematical model, it is possible to study the consequences of the requirements of invariance by mathematical derivations. (Andrich, 2004, p. 174)

With regard to the same point, Linacre (1996) argues that this is a virtue and not a flaw of the model.

Constant item parameters imply a constant construct. Different item parameters across samples of the relevant population imply that the construct has changed. Then measures cannot be compared across samples, and we are reduced to a vague notion of what we are measuring. (Linacre, 1996, p.513)

Furthermore, although invariance is a requirement of Rasch models, and of measurement, it is not an assumption for an analysis, in that one can test its veracity.

Criticism 5: Local independence

A different criticism refers to the assumption of local independence, which according to Goldstein (1979, p. 214) means that “for any individual, the response to an item is completely independent of his or her response to any other item”, again implying that this is not easy to find in practice.

Response to criticism 5

What the assumption means, in simple words (setting aside the statistical meaning) is that the response to any item should not affect the responses to other items. For example, previous items should not give hints, clues, insights or guidance for the solution of other items. Such an assumption is more like common sense, and can easily be met by experienced test constructors.

Athanasou and Lamprianou (2002), give an example of an item with sub-questions in simple arithmetic calculations.

“There are 18 flowers in John’s garden.

- (a) If he plants 6 flowers more, how many flowers will there be in total? Answer
- (b) If you need double the number of flowers, how many flowers will you need? Answer

These two parts of the item cannot be treated as different and independent. If a pupil is not in a position to find the answer to the first part, he/she will not find the answer to the second part even if he/she is able to double a number correctly.

Criticism 6: Symmetry between items difficulties and individual abilities

Goldstein (1979) also notes that the Rasch model “seems to imply a symmetry between item difficulties and individual abilities ... In reality, however, this is not the case” (p. 215)

Response to criticism 6

This appears to be a misunderstanding by Goldstein. The reference is presumably to the item-person map on Rasch software outputs. The Rasch model does not imply such

symmetry. However, the closer we are to such symmetry, the items are better targeted for the individuals, there is more information in the data and more accurate estimates (i.e. smaller standard errors) are obtained.

Criticism 7: Items need to be equally discriminating

Dickson and Kohler (1996) refer to the assumption that the Rasch model requires items to have equal discriminating power. An extension to that is Goldstein's (1979) argument that introducing a constant α_j in the model (discrimination parameter) makes the model more flexible and it is no longer necessary to have a constant relative difficulty between items. Although he acknowledges the increase in 'technical problems' in using the model with α_j , he states that "Because of its greater flexibility we can expect the model to have a better chance than model (3) (the Rasch model) of fitting a set of test scores." (Goldstein, 1979, p.215)

Response to criticism 7

To repeat: the aim of measurement models should not be to accommodate the fit of the test data but to satisfy the requirements of measurement. The aim is to measure, not to model. The 2-P model, which introduces a discrimination parameter, (and the 3-P model) seek to fit a model to the data not vice versa.

The Rasch model needs items to have discriminations that are equal enough to be regarded as the same. Misfit statistics act as quality control flagging items that fail to meet this measurement specification. In practice, according to Linacre (1996), unequal discrimination is diagnostic of various types of item malfunction and misinformation. Allowing or parameterising discrimination, which is a sample-dependent index limits the meaning of the measures to just that subset of items and persons producing this particular set of data. This prevents any general inferences over all possible items probing that construct among all possible relevant persons.

Criticism 8: The model is not perfect

Dickson and Kohler (1996) criticize the Rasch model in that no item fits the model exactly.

Response to criticism 8

The idea that the world is not perfect is not new. We use circles to approximate all sorts of round shapes and straight lines to describe objects that are not perfectly straight. If we were to stop investigations when things were not perfect we would do nothing.

A nice way of viewing the criticism is to take Andrich's (2004) line where he argues that the Rasch models, instead of simply describing data, provide the opportunity to understand data by the exposure of anomalies which is the prime function of measurement. The reason why the model can be used this way is that it formalizes conditions of invariance, which lead to properties of measurement. Thus, when the data deviate from the Rasch model it deviates from the requirements of measurement.

Similarly Linacre (1996) does not see non-fitting data as a criticism of the Rasch model but of the data. Failure of a data set to fit the Rasch model implies that the data do not support the construction of measures suitable for stable inferences. Linacre (1996) concludes that "usually, if the data have any meaning at all, they can be segmented into meaningful subsets that do fit the Rasch model and do support inferences" (p. 512), implying that even if the data are not unidimensional, when grouped appropriately (separating the dimensions) they will separately fit the Rasch model. The relevant question according to Linacre is not whether the items fit the model or not. It is 'Do the items fit the Rasch model well enough to construct useful measures?'

What any test constructor should be concerned with is that the basic assumptions of meaningful measurement should be satisfied. A test constructor with those assumptions in mind will construct test items that will yield data that will fit the Rasch model.

Criticism 9: All people do not fit the model

With regard to the persons' response patterns and whether meaningful inferences can be made from these response patterns, Dickson and Kohler (1996) comment that they have seen people who could climb stairs (considering them being successful on a difficult item) but not being able to swallow (considering them failing an easy item). The implied question in their argument is 'how can one make a meaningful inference from such a performance?'

Response to criticism 9

Again, when data do not fit the model they provide interesting anomalies to be investigated and to challenge the supposed scale. These anomalies are predicted by the Rasch model to occur occasionally, but are always unexpected when they do occur.

Finally Linacre (personal communication, September 5, 2006) quotes a paragraph from a New York Times Editorial stating:

That is the true test of a brilliant theory, says a member of the Nobel Economics committee. What first is thought to be wrong is later shown to be obvious. People see the world as they are trained to see it, and resist contrary explanations. That's what makes innovation unwelcome and discovery almost impossible.

An important scientific innovation rarely makes its way by gradually winning over and converting its opponents. ... What does happen is that its opponents gradually die out and that the growing generation is familiarised with the (new) idea from the beginning. No wonder that the most profound discoveries are often made by the young or the outsider, neither of whom has yet learned to ignore the obvious or live with the accepted wisdom.

“Naked Orthodoxy” (October 17, 1985)

Concluding remarks on the criticisms of the Rasch model

The Rasch model has turned the traditional relationship between data and model upside down. To consider blaming the data rather than the model when there is a mismatch between them is a considerable shift from the traditional, statistical way of thinking. The Rasch model however, was derived by Georg Rasch based on the property of invariance, not to describe any set of data but to provide in a mathematical and testable form the requirements of measurement. Most of the criticisms of the model have originated from this new approach to the data-model relationship.

Wright and Mok (2004) state that in order to construct inferences from observation a model with certain characteristics should be used. It must:

- Produce linear measures
- Overcome missing data
- Give estimates of precision
- Have devices of detecting misfit, and
- The parameters of the object being measured and of the measurement instrument must be separable.

Only the family of Rasch measurement models does this.

2.2.8 Validity and Reliability addressed through the Rasch model

2.2.8 (i) Validity

As it has been quoted earlier

Validity is an integrated evaluation judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores...what is evaluated is not the test but the inferences derived from the test scores.

(Messick 1993, 13)

Validity is a concept that can be addressed in part through the use of the Rasch measurement models.

If the items in a test or questionnaire are sufficiently well separated to define several statistically distinct levels, and hence a direction, we are ready to examine their ordering to see whether it makes sense. The pattern of item calibrations provides a description of the reach and hierarchy of the variable. This pattern can be compared with the intentions of the item writers to see if it confirms their expectations concerning the variable they wanted to measure.

To the extent that it does, it affirms the construct validity of the variable.

(Wright and Masters, 1982)



Items calibrated at much higher or lower positions on the variable than intended require further investigation for possible miskeying, short-cut solutions not noticed or unintended hints.

Wright and Masters (1982) argue that the internal validity of the test (i.e. whether the test items are consistent in measuring one variable) can be analyzed in terms of the statistical fit of each item to the model. They conclude that an item calibration is 'valid' if its mean square fit statistics are acceptable. Similarly if the mean square fit statistics of a person's performance are acceptable we can say that their measure is 'valid'. In other words the degree of the internal validity of a test or questionnaire is the extent to which the mean square fit statistics of the item calibrations and person measures are acceptable.

In a study of Callingham and Watson (2005), on measuring statistical literacy, item clusters were identified along the variable and a substantive interpretation of the underlying cognitive demands of the items within a cluster was undertaken revealing a series of levels along the variable that, taken together gave a description of the underlying construct. Furthermore, consistent fit to the Partial Credit Model of the data collected through the application of the test provided statistical evidence about the extent to which the separate items worked together to define a single construct. These two analyses provided evidence of validity against the conceptual and measurement model used.

Callingham and Watson (2005) state:

If the items are shown to be systematically and predictably related to each other along the variable (that is fit the model) this is confirmation that a single construct is being measured and provides evidence of construct validity. The extent to which test takers also fit the model provides further evidence that the test is behaving as intended. Consistent misfit of either items or persons' performance is a threat to construct validity (p. 23)

The extent to which a test measures one variable can be investigated further by factor analytic methods. Linacre (1998a) highlighted several options of factor analysis for identifying multidimensionality. These are factor analysis of (a) the observations, (b) the raw Rasch residuals, (c) the standardized residuals and (d) the logit residuals. He concluded that Principal Component Analysis (PCA) of the standardized residuals is the most effective in identifying multidimensionality.

Factor analysis of the original observations is informative of the factor structure but it does not construct the measures of the factors.

Also, the common logit scale, shared by person measures and item calibrations, “provides a picture of what a person can be expected to accomplish or endorse given the person’s ability and item calibrations (i.e., a criterion-referenced interpretation) within the boundaries of measurement error as quantified by the standard error” (Smith Jr., 2004a, p. 102).

Messick (1993, 1995) outlined the six facets of construct validity. Smith Jr. (2004a) argues that these facets may in part be addressed by the following three general aspects in Rasch measurement:

- i. The model requirements and measurement properties if the data fit the model
- ii. The order of items and persons on a common linear scale with the associated individual standard error and
- iii. The fit of the items and persons to the model requirements.

1. Content

Relevance and representativeness can be addressed through the rating by experts of the importance of each task/item. These ratings are calibrated to produce an order to the tasks/items on a linear scale from the most to the least important. Examining the empirical hierarchy and comparing it with the spread of the item calibrations along the variable provides an evaluation of the relevance and representativeness of the set of tasks/items.

The technical quality of items is addressed through item fit statistics. Misfitting items should be checked for possible technical faults.

2. Substantive

The substantive aspect of construct validity refers to theoretical rationales for the observed consistencies in test responses, ... along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks. (Messick, 1995, p.6)

According to Smith Jr. (2004a), these characteristics of construct validity may be addressed by verifying the definition of the variable intended by the researchers (confirmation of the intended item hierarchy) and examination of person fit statistics.

3. Structural

The structural aspect of construct validity addresses the credibility of the scoring structure to the structure domain.

The Rasch model has the following model requirements and measurement properties:

- The more able student should have a higher probability of answering any item correctly than a less able student and a more difficult item should have a lower probability of being answered correctly than a less difficult item, regardless of a person's ability.
- The cumulative total scores are sufficient statistics allowing for the separability of item and person estimates
- A raw score of any person (or item) represents the same amount of the variable being measured as the same raw score from a different person (or item).

If one believes these requirements are necessary for useful measurement, then the structural aspect of validity concerning how observations are combined into a score (sufficient statistics) and the scoring structure (how person ability and item difficulty must interact to govern the probability of an outcome ...) are satisfied.

(Smith Jr., 2004a, p.109)

4. Generalizability

The generalizability aspect of construct validity examines the extent to which score properties and interpretations generalize to and across population groups, settings and tasks.

This concept is stressed in the Rasch measurement literature through the property of invariance (Wright, 1967; Hambleton et al., 1991).

Smith Jr. (2004a) concludes that the generalizability of item and person measures depends on the fit of the data to the model and the invariance of parameter estimates over the classifications (e.g. time, groups and items) of interest.

5. External

Convergent evidence is sought through correspondence between different measures of the same, or related constructs, whereas discriminant evidence through the lack of correspondence between measures of distinct constructs.

Smith Jr. (2004a) claims that evidence for discriminant validity is sought through the Rasch model by a variation of the known Groups Method.

“Given two (or more) groups, that are hypothesized to differ in kind (not degree) on a variable, a researcher should be able to propose alternative item hierarchies for the two groups. To the degree that the empirical item hierarchies support the proposed item hierarchies, evidence of discriminant validity is obtained” (Smith Jr., 2004a, p.111).

As an example, Smith Jr. (2004a) describes a study in which Korean and American students were given an academic motivation scale. The interpretation of the results of that study led to the conclusion that for Korean students high academic motivation was driven by the importance of education as the means to social status (Statements like ‘It’s competitive and I like to compete’ and ‘Something that girls/boys are supposed to be good at’ were easily endorsed). On the other hand, for American students high academic motivation was driven by activities that they found satisfying (Statements like ‘I enjoy it’ and ‘It’s interesting to me’ were easily endorsed).

If these alternative hierarchies were proposed a priori, such evidence would provide support for external validity.

Smith's claim of a different approach does not seem like discriminant evidence of validity as described by Messick. Instead of this different approach one can always look for the lack of a relationship of the measures of the construct under investigation with measures of other distinct constructs.

With respect to convergent validity one can always investigate whether the scores from the instrument are related to scores from an already established instrument through the correlation coefficient.

6. Consequential

Rasch measurement does not directly address value implications of score interpretations and the potential consequences of test use.

However, fairness can be addressed through investigation of item bias (Smith, 1992). In Rasch measurement this means differences in item difficulties across the groups of interest. Furthermore the possible adverse impact of variations in judges' severity can be investigated by using the Many-Facet Rasch model developed by Linacre in 1989. For example, if two individuals of the same ability were rated by two judges, one lenient than the other. The individual rated by the more lenient would receive a higher raw score than the other individual. Using the Many-Facet Rasch model however would adjust the person measures taking into account the judges' severity estimates and provide a more valid and fair estimate of the individuals' abilities.

Also, the person fit statistics evaluate the believability of each person's response pattern and ability estimate and the associated standard error quantifies the precision of the estimate.

Detecting multidimensionality through Principal Component Analysis (PCA) of standardized residuals

The Rasch model uses the ordinal data to construct a one-dimensional measurement system regardless of the dimensionality of those data. Empirical data however, are always manifestations of more than one latent dimensions.

According to Linacre (1998a), the presence of more than one dimension in the data does not necessarily imply substantive multidimensionality. Extra dimensions may reflect different person response styles or different item content area. For example, items on subtraction may define a different dimension than items on addition in a simple mathematics test for young children.

Multidimensionality can also be an artifact of test construction. For example, including the identical item several times in a test produces a subset of highly intercorrelated items which may define an extra dimension. On the other hand, the use of different response mechanisms across items (multiple-choice, constructed-response, rating scales) introduces unmodeled variation which can be attributed to a dimension of 'item type'.

Multidimensionality only becomes a real concern when there are response patterns in the data indicating that the data represent two or more dimensions so disparate that it is no longer clear what latent dimension the Rasch dimension operationalizes.

(Linacre, 1998a, pp 5-6)

On a similar note, Smith Jr. (2004b) argues that unidimensionality should not be viewed as a dichotomous yes or no decision, but rather as a continuum. One has to decide at what point on this continuum multidimensionality threatens the interpretation of the item and person estimates.

Linacre (1998a) suggests that, for responses to complete tests, construction of Rasch measures from observational data, followed by PCA of Rasch standardized residuals provides the most effective means of identifying multidimensionality.

Linacre (2005) explains PCA of standardized residuals as it is used in WINSTEPS.

The purpose of PCA of residuals, according to Linacre (2005) is not to construct variables (as in factor analysis) but to explain variance in a possibly high-dimensional data set. First, one looks for the factor in the residuals that explains the most variance. If this factor is at the noise level, then the data is unidimensional as long as there is clear evidence for a scale, otherwise it is the second dimension, and then we look for a third etc.

Rotations are used in factor analysis to reapportion variance in an attempt to make the factor structure more interpretable, but, in doing so, the actual variance structure and dimensionality of the data are masked.

In PCA of the standardised residuals we do not want to find and interpret factors but to find the least number of factors above the noise level, explaining as much variance as possible.

The Rasch model is based on the specification of 'local independence'. This asserts that, after the contribution of the measures to the data has been removed, what is left is random, normally distributed noise. This implies that, when a residual is divided by its model standard deviation, the standardized residual of an observation is specified to be $N(0, 1)$ (Linacre, 1998a, 2005; Smith, 2000). Therefore the standardized residuals are modeled to have unit normal distributions which are independent and so uncorrelated. Consequently, all off-diagonal elements of the correlation matrix of the item standardised residuals are expected to be 0.

(The values put in the diagonal of the observed correlation matrix determine what proportions of the unit variances are factored into common factors. If 1s are placed in the diagonals, then principal components analysis results. That is, all the variance is explained by the components).

If we assert that all the variance in the standardized residuals is due to common factors and then put 1s in the diagonal we can test the assertion that the data conform to the Rasch model. If the resulting common factors explain nothing more than random noise across items, then the data conform to the Rasch model. The existence of substantive common factors, however, would indicate departure from unidimensionality.

“The aim of factor analysis of Rasch residuals is thus to attempt to extract the common factor that explains the most residual variance under the hypothesis that there is such a factor. If this factor is discovered to merely explain random noise, then there is no meaningful structure in the residuals.”

Linacre (1998b, p. 636)

Therefore a PCA of the standardized residuals identifies whether any other construct is shared in common among the items, i.e., presence of multidimensionality.

Procedure followed in PCA of the standardized residuals (Linacre, 2005, pp. 271 – 272)

1. The standardized residuals of all observations are computed.
2. Correlation matrices of standardized residuals across items (or persons) are computed.
3. In order to test the specification that the standardized residuals are uncorrelated, it is asserted that all randomness in the data is shared across the items and persons. This is done by placing 1s in the leading diagonal of the correlation matrix. This accords with the principal component approach to factor analysis.
4. The correlation matrix is decomposed. In principal if there are L items, then there are L item components. But these are expected to be random fluctuations in the structure of the randomness. However, an eigenvalue of less than 2 indicates that the implied dimension has less than the strength of 2 items, and so, however powerful it may be diagnostically, it has little strength in the data.
5. If items do have commonalities beyond those predicted by the Rasch model, then these may appear as shared fluctuations in their residuals. These will inflate the correlations between those items and result in components with eigenvalues greater than 1.
6. The total variance is expressed as the sum of cells along the leading diagonal, which is the number of items L. This corresponds to the unexplained variance in the dataset.
7. The variance explained by any factor is its eigenvalue.
8. Yardstick Power (YP) is the ratio of explained to unexplained variance in the dataset whereas the Power of the Yardstick relative to a specific factor (YF) is given by:

$$YF = YP \cdot \frac{L}{\text{eigenvalue}} .$$

A key issue in the interpretation of PCA is the choice of the critical value of the eigenvalue. Smith and Miao, according to Raiche (2005) and Linacre (2005), used simulated data and indicated that eigenvalues less than 1.4 are at the random level, whereas Smith Jr. (2004a) decided, by using three sets of simulated data, that an eigenvalue greater than 1.5 (in a 30 item instrument) would be considered as representing the existence of a second dimension.

Raiche (2005) simulated data for various numbers of items and subjects and reported that 1.4 was always exceeded by the first and usually second eigenvalue. His recommendation is to decide the criterion eigenvalue directly from relevant simulations.

Linacre (2005) in his description of PCA of the standardized residuals gives an example, where the eigenvalue of the first factor extracted was 2.7 (14 items were used). Although it seems like a high value, indicating the presence of a second dimension, its strength is very small (it explains only 0.2% of the total variance in the data and it is about 560 times smaller than the variance explained by the dimension measured by the test). Linacre implies with this example that perhaps more importance must be placed on the strength of the factors and not on the magnitude of their eigenvalues.

In concluding, and having in mind what he was implying with the strength of the factor, he gives some general rules of thumb, one concerning the eigenvalues, is that in the unexplained variance a secondary dimension must have the strength of at least 3 items. If a factor has eigenvalue less than 3 (in a reasonable length test) then the test is probably unidimensional.

But perhaps a more effective way of detecting multidimensionality is the use of loadings against item locations plots. Linacre (1998a) compared factor analyses results from the observational data, the raw, standardized and logit residuals through plots of the factor loadings against item difficulty calibrations. In such plots items located on different dimensions will be seen to cluster together. He concluded that PCA of standardized residuals is the best method for detecting the presence of more than one dimensions.

2.2.8 (ii) *Reliability*

Using the Rasch model provides a direct measure of the test error variance S_{ϵ}^2 which tells us how precisely one will be able to estimate each person's ability when the items are internally consistent. The estimate of the standard error is not influenced by sample variance or fit and so it is not sample specific. It is a sample-free test characteristic of the set of items, which make up the test. It estimates how precisely the ability of each person whose response pattern fits can be estimated from their particular score on the test, regardless of any sample to which he may belong. Unlike the traditional reliability coefficient and the measurement error it implies, this estimate is not an average for the whole test but is particular to the test score the person actually obtains.

(Wright and Masters, 1982, p. 113-114)

Therefore the great advantage of reliability estimated when using the Rasch model is that the estimate of the standard error is specific for each person, based on his test score and is not group dependent.

Two important reliability indices are reported in Rasch analyses.

The Person Estimate Reliability is an indication of the precision of the instrument and shows how well the instrument can distinguish individuals. According to Curtis (2004), Andrich (1982) has shown that this index is virtually identical to the KR-20 or its generalization Cronbach's alpha. Linacre (1999) also relates the Rasch person separation reliability with Cronbach's alpha. Both of these are estimates of the ratio of "true measure variance" to "observed measure variance". The basic underlying relationship is specified to be:

$$\text{Observed Variance} = \text{True Variance} + \text{Error Variance}.$$

The Item Estimate Reliability shows how well the items that form the scale are discriminated by the sample of respondents. Wright and Masters (1982, pp 90-92) argued that good item separation is a necessary condition for effective measurement.

Smith Jr. (2004b) refers to the following problems in using the KR20 formula as a measure of internal consistency:

The 'average' person variance used in KR-20 will always overestimate the error score variance of persons with high or low scores (since persons with high or low scores have less error variance than persons with scores near 50%).

Also in many studies, estimates of internal consistency are reported based on previous applications of an assessment and these are not informative unless the proposed sample has exactly the same score distribution as the sample used for the reported internal consistency.

Furthermore the use of raw scores in calculating the sample variance is probably misleading since raw scores are not linear. The reliability estimate is then used in the calculation of the standard error of measurement, which in turn is used to represent the precision of every possible score on the scale, even though it is known that extreme scores are less precise than central scores.

Linacre (1999) refers to another problem with Cronbach's alpha which explains also why alpha is usually higher than the Rasch separation reliability. In the calculation of Cronbach's alpha extreme scores (full marks or zero marks) are included. Since these extreme scores have no score error variance, their effect is to increase the reported reliability.

Furthermore, Cronbach's alpha is also computed from non-linear raw scores.

However the Rasch separation reliability for N examinees is computed from linear measures by:

$$R_p = 1 - \frac{\sum (\text{Measure Standard Error})^2 / N}{\text{Variance of Observed Measures}}$$

These correlational-based reliability estimates (like KR20, Cronbach's alpha and R_p) are non-linear. For example an improvement in alpha or R_p from 0.5 to 0.7 is not twice the improvement from 0.85 to 0.95. Furthermore these estimates of reliability suffer from the restricted range of values they can take, that is, from 0 to 1.

According to Wright and Masters (1982) R_p can often be replaced by G_p , a person separation index which ranges from 0 to infinity and is calculated by:

$$G_p = \sqrt{\frac{R_p}{1 - R_p}}.$$

G_p is on a ratio scale and compares the true spread of the person measures with the measurement error and indicates the spread of person measures in standard error units. Therefore the higher the value of G_p , the more spread out the persons are on the variable being measured.

Another useful calculation is that of strata.

$$strata = (4G_p + 1)/3.$$

Strata are used to determine the number of statistically distinct levels, separated by at least 3 errors of measurement, of person ability that the items have distinguished (Wright and Masters, 1982)

2.3 Appropriateness Measurement

2.3.1 Introduction

Appropriateness measurement (AM) is concerned with the investigation of individual score patterns and in particular the unusual, aberrant or inappropriate score patterns. An aberrant score pattern is one that is improbable, given either that an IRT model fitted the data or given the item score patterns of other persons in the group. Drasgow, Levine and Williams (1985) define AM as

“a model-based attempt to control test pathologies by recognizing unusual patterns”.

If an aberrant response pattern is discovered during the test, and this is possible in computerized adaptive testing, then this is evidence that the person is taking the test inappropriately and the test may be halted and the reasons for the aberrance can be directly investigated. If however it is discovered following the test, the inferences from the test score may be withheld until further investigation.

The study of aberrant scores 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).

Possible sources of aberrant behaviour include cheating, sleeping or carelessness, guessing, alignment errors, plodding and item bias (Karabatsos, 2000; Meijer, 1996; Rudner, 1983; Wright, 1977). Other possible sources are test anxiety (Harnisch and Linn, 1981; Athanasou and Lamprianou, 2002), copying, sudden illness and special knowledge (Linacre and Wright, 1994), low language fluency (Rudner, 1983) and item multidimensionality, misworded items, disordered papers in test booklets or miskeyed items (Karabatsos, 2000). Furthermore, Tatsuoka and Tatsuoka (1982) offer empirical evidence that patterns of aberrant responses relate to differences in instruction.

Karabatsos (2000) groups the measurement disturbances within educational testing into three different levels.

At the examinee level

An unexpected series of correct responses to difficult items may indicate *cheating*, whereas a few unexpected responses *lucky guessing*. On the other hand, a series of unexpected incorrect responses to easy items could be an indication of *deficient sub-abilities* whereas, a few unexpected incorrect responses of *sleeping or carelessness*. *Random guessing* or *extreme creativity* could lead to unexpected correct responses to hard items and at the same time unexpected incorrect responses to easy items.

At the item level

Item multidimensionality (when a subset of items do not measure the same attributes as the other items) could lead to measurement disturbances and so can *item bias* (i.e. when a certain examinee group responds differently to an item than another group). *Multiple correct response options* for an item could lead to confusion and unexpectedly correct or incorrect responses and *misworded items* can cause examinees to misinterpret that item.

At the test administration process

Disordered pages in a test booklet and *miskeyed items* can also lead to confusion amongst examinees and to measurement disturbances.

These threats to the examinee measurement accuracy occur too often in various test administrations. Therefore appropriateness measurement methods employed should be able to detect aberrant responses in a highly reliable and accurate fashion.

Measurement disturbances can also threat attitude measurement.

Curtis (2004) mentions social desirability, acquiescence, self-awareness, irrationality, inadmissibility, self-incrimination and politeness as such disturbances. These may lead to reduced precision in item and scale parameters and may influence the fit of persons to the instruments.

Attitude survey instruments, in contrast to achievement tests, are rarely high stakes activities and for this reason, some participants may respond carelessly and therefore compromise the calibration of the instrument. It is also well known that some people fall into an inappropriate pattern of responses, such as checking all items on the right

hand side, hence the advice is to word items in such a way that respondents will be required to vary the position of their ticks to give consistent responses.

2.3.2 Possible Factors associated with misfit

Gender

Much research has been carried out on whether gender affects performance on achievement tests. For example, Plake et al. (1982) reported that for mathematics achievement tests, with highly motivated students taking part, the sex of the subject interacts with the item arrangement yielding significantly higher scores for males more on easy-hard ordering than under any other item arrangement. They also argue that their findings are in accordance with similar researches documenting male superiority on such tests, like the ones by Fennemna and Sherman in 1974 and Benbow and Stanley in 1980.

On the effect of gender on aberrance, Frary and Giles (1980) showed that overall whites and females had lower person fit statistics values, indicating lower aberrance for these two groups, as opposed to blacks and males.

Item arrangement

According to Plake et al. (1982) item arrangement appears to be an important variable that can, in fact, influence test performance. The male superiority in mathematics achievement tests was more significant in an easy-hard ordering. Perception scores (difficulty and performance) are also influenced by item ordering. It is well established that, when tests are speeded, the easy-to-hard ordering of the items is best from a psychometric perspective. Towle and Merrill (1975) state that although Sax and Cronbach, in 1966, supported the advisability of easy-to-hard sequencing of items when testing time is severely restricted. They concluded that little is gained from arranging test items in ascending difficulty, if time limits are generous or non-existent. Towle and Merrill (1975) suggested that items in a timed test could be arranged in a random or easy-to-hard order but not in a hard-to-easy order since performance is impaired.

Mismatch between curriculum and test content

Harnisch and Linn (1981) studied the effect of school and regional differences on the caution index and concluded that schools in different parts of the state had very different indices. The sample used in their study consisted of 110 schools and 6300 students (approximately 2100 from each of grade levels 4, 8 and 11).

They attributed this school effect to the fact that certain schools may have not covered segments of the content sampled by the test, or that they have given less emphasis to some of the content. In other words their suggestion was that the differences in the index were caused by a mismatch between school curriculum and test content.

Test anxiety

It is well known that test anxiety generally relates to test performance. The strength of this relationship depends to a large extent on the perceived importance of the testing situation (Sarason and Palola, 1960). O'Reily and Wightman (1971) extend the findings of other authors like Hill and Sarason that there is a negative relationship between anxiety and achievement test performance, by arguing that in research where the negative relationship is non-existent, one of the major reasons is the tendency of some children to lie about their anxious feelings, to be defensive thus depressing their true scores on questionnaires measuring anxiety.

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

According to Bracey and Rudner (1992), Schmitt and Crocker investigated the relationship between scores on the Test anxiety scale for adolescents and person-fit.

They reported that students in the middle ability range showed no relationship between test anxiety and person fit indices.

High-ability, low-anxiety students showed greater misfit than high-ability, high-anxiety students whereas at the low-ability end the reverse was true; low-ability, high-anxiety students showed greater misfit than low-ability, low-anxiety students.

Position on the ability/trait scale

Masters and Keeves (1999) expressed concerns about trait range affecting misfit, suggesting that persons in different ability ranges could have different proportions of misfits. However, Curtis (2004) makes reference to Li and Olejnik (1997), who compared the performances of five misfit indicators and found no correlation between trait estimate and misfit with any of the indicators. This, according to Curtis, suggests that the concern expressed by Masters and Keeves are not a matter of great concern.

On the other hand, Petridou and Williams (2007) report that high ability students can manifest more aberrance and this can be attributed (as explained by the pupils themselves in interviews) to carelessness and silly mistakes.

Attention Deficit Hyperactivity Disorder (ADHD)

ADHD is a specific developmental disorder that comprises deficits in behavioral inhibition, sustained attention and resistance to distraction, and the regulation of one's activity level to the demands of a situation.

According to Barkley and Murphy (1998), since 1980, it has become possible to place those with ADHD into subtypes depending on the symptoms they experience. Those who are diagnosed as have particular difficulties primarily with impulsive and hyperactive behavior and not with attention or concentration are referred to as having *ADHD, Predominantly Hyperactive-Impulsive Type*. Individuals with significant inattentiveness, without being impulsive or hyperactive are called *ADHD, Predominantly Inattentive Type*. However, most individuals with the disorder will manifest both of these clinical features and thus are referred to as *ADHD, Combined Type*.

The Diagnostic and Statistical Manual of Mental Disorders Version 4, (DSM-IV) developed by the American Psychiatric Association in 1994, contains a list of 18 criteria

for the diagnosis of ADHD. The guidelines specify that for the children to be diagnosed as having ADHD, they must meet at least 6 out of the 9 criteria relating to inattention for the Predominantly Inattentive subtype and at least 6 out of the 9 criteria relating to hyperactivity and impulsivity for the Predominantly Hyperactive/Impulsive subtype. For the Combined subtype they must meet both of the above conditions.

Barkley and Murphy (1998, pp. 6-7) report that ADHD occurs in approximately 3 – 7 % of the childhood population in the USA, with a ratio of boys to girls of 3:1 and approximately 2 – 5 % of the adult population with a ratio of males to females of 2:1.

However, Merrell and Tymms (2001) estimated the proportion of children observed by their teachers to display severe ADHD symptoms in the UK to be higher, between 8.1 % and 17 %.

Furthermore, Merrell and Tymms (2005) reported that inattentiveness was more associated with a negative impact on academic progress than hyperactivity/impulsivity.

On the effect of ADHD on the reasonableness of the response patterns an investigation into the possible association of ADHD behaviour and misfit was carried out at the CEM centre of the Durham University by Panayides, Merrell and Tymms (2008). They found no relationship between ADHD, gender and misfit for the test comprising of only constructed-response items, but highly significant links in the test comprising of only multiple choice-items. Although boys with and without ADHD symptoms had similar proportions of misfit, girls with ADHD symptoms had significantly higher proportions of misfit than girls without. The combination of gender, ADHD symptoms and type of test items had a significant effect on misfit. Girls with ADHD symptoms had a much higher proportion of misfits in multiple-choice mathematics items.

Mathematics Self-Concept

Academic self-concept is defined as the general feeling of doing well or poorly in school.

Shavelson, Hubner and Stanton (1976) argued that self-concept is a multifaceted hierarchical construct and that in particular self-concept in different academic areas combine to form a higher order academic self-concept. Their argument, according to Marsh, Byrne and Shavelson, (1988), was based partly on conceptually similar models

of ability that posit a higher order ability factor as well as more specific components of ability.

Marsh and Shavelson (1985) found no significant correlation between mathematics and verbal self-concept and those did not combine with school self-concept to form a single second order academic factor.

Marsh developed in 1986 the Internal/External frame of reference model to account for the extreme separation of math and verbal self-concepts and their relations to math and verbal achievement. He showed that math and verbal achievement correlate higher with the matching areas of self-concept than with the general academic self-concept.

In terms of gender differences in math self-concept, many researchers (such as Marsh et. al (1988); Skaalvik and Skaalvik, (2004)) found that male students had higher self-concept, meaning that males seem to judge themselves more favourably than females do, as early as the end of elementary school. However, none of the gender differences in maths self-concept could be explained by differences in achievement. This supports the gender stereotype explanation of gender differences in self-concept and motivation, which predicts that the gender differences in self-concept are larger than can be explained by the differences in achievement.

Motivation

Lamprianou and Boyle (2004) argue 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.

Class effect

Petridou and Williams (2007) report a high class level effect on aberrance. They suggest the following reasons for this significant effect:

- Non-standard administration practices such as teachers interpreting questions.
- Class 'cheating' (p. 243) by leaving materials related to the test on the classroom walls.

- Instructional effects in terms of topics not being taught by the time of the test administration.

Identifying aberrant responses using a test data matrix

Aberrant responses and possible sources of measurement disturbances can be identified using a test data matrix. The table 2.3 below shows a test data matrix containing the responses of 20 students to 10 multiple-choice items in algebra. It is composed of 0s (for incorrect responses) and 1s (for correct responses). There are 20 rows, one for each student and 10 columns, one for each item.

Each row contains the responses of one student to the 10 items in the test. A number on the left of the matrix identifies each student. By summing across a student's row of responses, a score is obtained for that student. The 20 students have been sorted in descending order, by score, from top to bottom.

Each column contains the responses of the 20 students to one item. The entries in each column are summed down the matrix over the 20 students to obtain a score for that item. The 10 items have been sorted so that the easiest item is on the left of the matrix and the rest follow in increasing difficulty, with the hardest item being on the right of the matrix.

Table 2.3: Test data matrix

Examinees	item1	item4	item3	item2	item5	item8	item7	item6	item9	item10	score
1	1	1	1	1	1	1	1	1	1	1	10
10	0	1	1	1	1	1	1	1	1	1	9
16	1	1	1	1	1	1	1	1	0	1	9
2	1	1	1	1	1	0	1	0	1	1	8
14	1	1	1	1	1	1	1	1	0	0	8
9	1	1	1	1	1	1	0	1	0	0	7
11	0	1	1	0	1	1	1	1	1	0	7
15	1	1	1	0	1	0	1	1	0	1	7
13	1	1	1	1	1	1	0	0	0	0	6
4	1	1	1	1	1	0	0	0	0	0	5
5	1	1	0	1	1	0	1	0	0	0	5
12	1	1	1	1	1	0	0	0	0	0	5
6	1	0	1	1	0	0	0	0	1	0	4
7	1	0	0	0	0	1	0	0	1	1	4
17	1	1	1	1	0	0	0	0	0	0	4
19	1	1	1	1	0	0	0	0	0	0	4
18	1	1	1	0	0	0	0	0	0	0	3
20	1	0	0	1	0	0	1	0	0	0	3
3	1	0	0	0	1	0	0	0	0	0	2
8	0	1	0	0	0	1	0	0	0	0	2
Item Score	17	16	15	14	13	9	9	7	6	6	

The 1s are expected to pile up on the top left of the matrix (where we have the easiest items and the students with the highest scores) and the 0s in the bottom right (where we have the hardest items and the students with the lowest scores).

A row of misplaced 1s or 0s is a sign that a student has performed in an unusual way.

Students 10 and 16, for example, both scored two of the highest scores in the group. However student 16 failed on one of the hardest items (as could be expected) whereas

student 10 failed the easiest item in the test. This probably means that the second student made a careless mistake.

Students 6, 7, 17 and 19 are all low scorers with a total of 4 correct responses out of the 10 items.

Students 17 and 19 responded exactly as expected by examinees of their ability (their response pattern is perhaps too good to be true). It could however be plodding behavior by slow and methodical examinees who refuse to proceed to the next question until they have done their utmost to answer the present item correctly. On the other hand, student 6 has responded unexpectedly correctly to one of the most difficult items, probably by lucky guessing, and student 7 answered correctly the two hardest items and that could be an indication that he may have copied the answers from a more able neighbour. Student 13 is another student whose response pattern may be too good to be true. He may be a plodder too.

Close inspection of the test data matrix could help identify possible aberrant responses however it only gives an indication as to possible reasons for the aberrant patterns.

Many authors (such as Meijer, 1996; Molenaar and Hoijtink, 1996; Athanasou and Lamprianou, 2002) agree that after identifying misfitting examinees, further qualitative investigations concentrating on the examinees, such as interviews could reveal the real reasons for the aberrant response behavior.

Extensive research in the second half of the 20th century produced a body of appropriateness statistics. Those statistics are commonly known as grouped-based indices because they study the agreement of individual responses with the responses of the rest of the group aiming to identify unexpected response patterns, which could lead to invalid measures of examinees' ability.

According to Meijer and Sijtsma (1999) several of these indices usually counted certain score patterns for item pairs and compared this count to the expectation of the Guttman model, which assumes that any examinee who gives a correct response to a difficult item must also give correct responses to easier items or any examinee who gives an incorrect response to an easy item should respond incorrectly to the more difficult items.

2.3.3 Person-Fit Statistics

Appropriateness indices were popular in the late 1970s, however, the probabilistic nature of the IRT models proved to be an attractive basis for the development of a new series of indices. These indices are usually called 'person-fit statistics' because they mainly evaluate the fit of an IRT model to the response patterns of examinees. Frary (1982) describes person-fit as "the extent to which an examinee's response pattern ... is consistent with his ability as estimated by total score" (para. 1)

Person-fit statistics are measures of the degree of reasonableness, or 'indicators of the believability' (Smith, 1986), of examinees' answers to a set of items. They inform the researcher of the extent to which an examinee has responded to the items in ways consistent with the other examinees in the sample. A large person fit implies that the person's pattern of responses is not consistent with that predicted by the model.

Therefore person-fit statistics are important in detecting aberrant response patterns that could lead to inaccurate measurement.

Curtis (2004) reports the following:

The inclusion of responses that underfit the Rasch measurement model, ... increase the standard errors of the item estimates, reduce the range of item locations on the scale, and reduce the inter-threshold range within items. Thus, the inclusion of misfitting cases compromises the measurement properties of the scale formed by the instrument (p. 141)

Emons et al. (2003) mention the following uses of person fit analysis.

- It can be used to identify misfitting students so as to be reassessed by another test in order to obtain a more valid estimate of their ability.
- In the context of education, person misfit may lead to the decision of remedial teaching of certain abilities and skills so as to have more valid test performances.
- At the test administration level, results from person fit analysis may help to improve test conditions.

- At the data analysis context, misfitting item score vectors may be considered to be outliers and data analysis may compare the results obtained from the complete data, including the outliers, and the data without the outliers.

In the literature on fit indices, there has been considerable emphasis on item fit and even in introductory books (like Bond and Fox, 2001, pp. 179-183) the emphasis is on interpretations of fit indices for items. Wright (1995) quoted Rudner et al. who claimed that the research on person fit statistics has been largely unsystematic, atheoretical and not been explored in applied settings.

Curtis (2004) comments on this criticism by saying that it appears to be harsh, as a considerable body of work has emerged since the late 1980s.

However, in most studies of fit indices, dichotomous test data have been the main concern. (Curtis 2004, p. 126; Karabatsos 2000, p. 170).

Curtis (2004) then adds that attitude instruments warrant specific attention mainly because they are rarely high stakes instruments and so respondents' behavior may be rather different from test behavior and the number of response categories may interact with misfit indicators.

The most important person-fit statistics can be categorized to the following groups. (They are briefly described in Lamprianou, 2002)

The first group consists of the residual-based person-fit statistics (Karabatsos, 2000). These statistics aggregate discrepancies between the expected responses of the examinees and their actual responses. Typical representatives of these are the Infit and Outfit mean square statistics (Wright, 1977; Wright and Masters, 1982).

The second group of person-fit statistics (the likelihood-based statistics) consists of those indices based on likelihood. A major representative of this category is the ℓ -statistic presented in 1979 by Levine and Rubin. This statistic is the log-likelihood of an examinee with ability θ to generate a particular response pattern. Drasgow, Levine and Williams (1985) put forward a standardized version of ℓ , and named it ℓ_z .

The third group consists of the family of Caution indices. Sato proposed his Caution index in 1975 (presented in Harnisch and Linn, 1981). This index was used to indicate that caution is needed in interpreting response patterns that were flagged as aberrant.

Sato used a data matrix of examinees responses (0s and 1s) in the rows, with the highest scoring examinees on the top and the lowest scoring examinees at the bottom. At the same time item responses were put in the columns, from easiest to hardest from left to right. This matrix has been called Student-Problem (S-P) Table.

If the items formed a perfect Guttman scale the S-P table would consist of a section with all ones in the upper left-hand corner and all zeros in the bottom right-hand corner. In practice, perfect Guttman scores cannot be expected on achievement test items, consequently the S-P table will contain a vast majority of ones in the upper left-hand corner and a vast majority of zeros in the lower right-hand corner.

Sato constructed two step-lines on the table. Using the examinees' total score (number of correct responses) he drew the first step line (the S-curve) by constructing a perpendicular line in each row such that the number on cells on the left of this line is equal to the score of that examinee.

The second step line (P-curve) was drawn in a similar fashion using the item scores (the number of examinees responding correctly to an item). A horizontal line was drawn in each column such that the number of cells above that line was equal to the score on that item. (See table 2.4)

Table 2.4: The S-P table

	item 1	item 2	item 6	item 5	item 3	item 7	item 4	item 8	Score
examinee 1	1	1	1	1	1	1	1	1	8
examinee 12	1	1	1	1	1	1	0	1	7
examinee 4	1	1	0	1	1	1	1	0	6
examinee 7	1	1	1	1	1	1	0	0	6
examinee 3	1	1	1	1	0	0	0	0	4
examinee 5	1	1	1	0	1	0	0	0	4
examinee 11	1	0	1	1	0	0	1	0	4
examinee 2	0	1	0	1	1	0	0	0	3
examinee 10	1	1	1	0	0	0	0	0	3
examinee 15	1	1	0	0	0	0	1	0	3
examinee 6	1	0	0	0	0	1	0	0	2
examinee 9	1	0	1	0	0	0	0	0	2
examinee 13	0	1	0	1	0	0	0	0	2
examinee 8	1	0	1	0	0	0	0	0	2
examinee 14	0	0	0	0	1	0	0	0	1
Score	12	10	9	8	7	5	4	2	

Ideally the S- and P- curves should coincide. The index was based on the area between the two curves, which is potentially useful in evaluating the homogeneity of the test.

The key point is that the caution index provides information about an examinee that is not contained in the total score. A large value raises doubts about the validity of the interpretation of the total score of an individual.

The final group (like the third group) consists of non-parametric person-fit statistics. Non-parametric person-fit statistics are calculated given that a non-parametric IRT model fits the data or given the score patterns of the other examinees in the group. U_3 is a typical representative of these person-fit statistics and was developed by Van der Flier in 1982.

According to Lamprianou (2002):

Although non-parametric person-fit statistics are very promising (they can be used in the context of non-parametric Item Response models which are very useful when only ordinal data are available), they have not yet been extensively studied and applied. It has been shown (Meijer, Muijtjens and Van der Vleuten, 1995) that under certain conditions they can have a similar detection rate with the group-based indices. (p. 49)

Karabatsos (2003) gives a table of 36 person-fit statistics, 11 non-parametric and 25 parametric, and a brief description of those, together with 11 more making a total of 47 statistics. The large number of these statistics found in the literature makes it difficult for a researcher to decide which one to use in practical situations.

Molenaar and Hoijsink (1996, p. 28) suggest the following:

- Use a person-fit statistic whose distribution under the null hypothesis of model conformity is known or at least roughly known (Molenaar and Hoijsink, 1990).
- When a particular aberrance is suspected use a statistic that has power against it (Klauser, 1995).
- Otherwise use a statistic that has at least some power against the most serious types of aberrance.

2.3.4 Infit and Outfit mean square statistics

2.3.4.(i) Introduction

All empirical data departs from the Rasch model to some extent. How much of this departure is tolerable?

In regression analysis fit statistics are used to discover a model that fits the data well enough so as to consider that it generated the data.

In Rasch analysis the model is already chosen. The purpose of the fit statistics is to aid in measurement quality control, to identify those parts of the data which meet Rasch model specifications, and those parts which don't. Parts that don't are not automatically rejected, but are examined to identify in what way, and why, they fall short, and whether, on balance, they contribute to or corrupt measurement. Then the decision is made to accept, reject or modify the data.

(Smith, 1996, p.516)

Infit and outfit when using the dichotomous Rasch model

These statistics were first introduced by Wright and Panchapakesan (1969), who developed the first fit statistic, the overall Chi square statistic, used to assess the fit of the entire data matrix to the Rasch model and also demonstrated the use of the item fit statistic. The outfit and infit were further elaborated by Wright (1977) and Wright and Masters (1982).

Outfit is based on the conventional sum of squared standardized residuals. Linacre and Wright (1994) describe how these statistics can be calculated.

If X_i is the observed score on item i , E_i is its expected value (which for the dichotomous model is equal to p_i , the probability of answering an item correctly), based on the parameter estimates and σ_i^2 is the modeled variance about this expectation, then the

squared standard residual is given by: $Z_i^2 = \left(\frac{X_i - E_i}{\sigma_i} \right)^2$ and the outfit mean square statistic by:

$$\text{Outfit} = \frac{\sum_{i=1}^N Z_i^2}{N} \quad \text{where } N \text{ is the number of observations summed.}$$

The outfit statistic 'is dominated by unexpected outlying, off-target, low information responses and is outlier-sensitive' (Linacre and Wright, 1994).

To reduce the influence of outliers a weighted mean square can be calculated by weighting Z_i^2 by the information available. The statistical information in a Rasch observation is its variance, which is larger for targeted observations and smaller for extreme observations.

Therefore, infit is an information-weighted sum.

$$\text{Infit is given by:} \quad \text{infit} = \frac{\sum_{i=1}^N Z_i^2 \cdot \sigma_i^2}{\sum_{i=1}^N \sigma_i^2} = \frac{\sum_{i=1}^N (X_i - p_i)^2}{\sum_{i=1}^N \sigma_i^2}$$

'Infit is dominated by unexpected inlying patterns among informative, on-target observations and so is inlier-sensitive' (Linacre and Wright, 1994).

Linacre (2006) explains that in the Rasch context, outliers are often lucky guesses or careless mistakes, which can make a 'good' item look 'bad'. Consequently, infit was devised as a statistic that downweights outliers and focuses more on the response string close to the item difficulty (or person ability).

In answering a question about which of the two mean squares should be reported, Linacre (2006) recommends reporting the outfit because:

- It is easier to interpret
- Statisticians are familiar with it (being a conventional Chi-square)

The recommendation about infit is to avoid reporting it (because it is more difficult to diagnose and interpret, and it is also unfamiliar to statisticians), unless the data are heavily contaminated with irrelevant outliers.

Infit and outfit when using the Partial Credit Rasch model

Masters and Wright (1997) give a description of the outfit and infit statistics and how they are used when the Partial Credit Model is applied, using a slightly different notation.

For person j , with ability θ_j , item i , the person score $x_{ij} \in [0, 1, \dots, m_i]$ has expectation

$$E_{ij} = \sum_{k=0}^{m_i} k P_{ijk} ,$$

where P_{ijk} is the probability of person j scoring K on item i ,

and variance
$$W_{ij} = \sum_{k=0}^{m_i} (k - E_{ij})^2 P_{ijk} .$$

Then the residuals are given by $y_{ij} = x_{ij} - E_{ij} .$

A positive residual indicates that the observed score is higher than that expected whereas, a negative residual indicates that the observed score is lower than that expected.

The standardized residuals are given by:
$$z_{ij} = \frac{x_{ij} - E_{ij}}{\sqrt{W_{ij}}} .$$

The Outfit statistic for each person is the mean of the squared standardized residuals over all items. That is:

$$u_j = \frac{1}{n} \sum_{i=1}^n z_{ij}^2 .$$

Infit is the sum of the squared residuals over all items divided by the sum of the variances of all observations.

$$\text{That is } v_j = \frac{\sum_{i=1}^n w_{ij} z_{ij}^2}{\sum_{i=1}^n w_{ij}} = \frac{\sum (x_{ij} - E_{ij})^2}{\sum_{i=1}^n w_{ij}} .$$

2.3.4.(ii) Critical values for the infit and outfit mean square statistics

Wright et al. (1994) provide a table of reasonable mean square fit values and suggest item 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 1.3 indicate 30% more variability and values of 0.7 indicate 30% less variability than predicted by the Rasch model. In such a case, a person’s response pattern with infit or outfit statistic above 1.3 is considered unexpected or unpredictable (misfit) and below 0.7 too predictable, and flagged as overfit.

Overfit is usually ignored as it is not considered a disturbance to measurement. It simply means that the specific response pattern is too close to a Guttman response pattern. That is, the examinee answers correctly questions with difficulty lower than his/her ability more frequently than expected by the Rasch model. Also it means that the examinee answers incorrectly questions with difficulty higher than his/her ability more frequently than expected by the model.

Linacre and Wright (1994) explain why such a response pattern is flagged as problematic and not considered ideal. They say that it is like splitting the test into two subtests, an easy test on which the person performed infinitely well and a hard test on which the same person performed infinitely badly. This increases the uncertainty in the reported measure and raises the question whether the sharp transition is really a precise indicator of the person’s measure or whether it was caused by other factors such as time limits, response style, curriculum effect or sudden illness.

Keeves and Alagumalai (1999) comment that it is customary for items to be considered to fit the Rasch model if they have item infit or outfit mean square statistic in the range 0.77 to 1.30, although many researchers would prefer to use the more restricted range from 0.83 to 1.20.

They also suggest that for small samples and short tests, a correction should be applied to the values of the infit and the outfit, using correction factors of $L/L-1$ and $N/N-1$ to allow for bias, where L is the number of items and N the number of persons.

Bond and Fox (2001, pp 177-183) suggested ranges of acceptable fit statistics too, for various test and survey instruments and provide some discussion of the meanings that might be attached to misfit. Curtis (2004, p.141) reports that instrument targeting or mis-targeting, item and person variance, instrument length and the number of response options all influence the distribution of the infit and outfit mean squares. His findings suggest that it is possible to provide only broad guidelines about the critical values that might be used to discriminate fitting from misfitting cases. He suggests as an acceptable range for the infit and outfit for the two attitude instruments he examined from 0.5 to 1.6, quite close to Bond and Fox (2001, p. 179) who suggested 0.6 to 1.5.

The reason for using a wider range of acceptable fit statistics for attitude instruments or personality scales is that the more control there is over the testing situation the tighter fit we can demand. Linacre (personal communication, March 7, 2007) states:

“For high stakes multiple-choice tests the items are highly controlled, carefully constructed and piloted and the examinees respond in a highly controlled environment. Questionnaires are usually less carefully constructed and there is less control over how respondents behave. Observational instruments usually have even less control (or even no control) of how respondents behave.” Linacre (personal communication, March 7, 2007) concludes by stating:

“less control $\Rightarrow$ more off-dimensional behaviour $\Rightarrow$ worse fit expected”

The primary purpose of conducting a test is to measure the ability of examinees. One needs measures that are appropriate for his/her purposes. Rough measures are useful for the purposes of assessing personality traits therefore the fit criteria can be much more relaxed. Rough measures are probably useful enough for classroom teachers too, therefore the fit criteria can also be more relaxed.

However, when certifying the competence of a medical practitioner, or when students take university entrance exams in a highly competitive environment rough measures are not good enough, therefore much tighter fit criteria are applied.

Curtis (2004) recommends using simulation studies to establish critical values for the fit statistics separately for each instrument used. Also Glas and Meijer (2003) suggest using simulated data according to an IRT model based on the estimated item parameters and then determine the critical values empirically.

Although researchers have proposed various cut-off scores for identifying misfit, these are just rules-of-thumb. One should always check the data carefully and thereby apply different cut-off scores. Especially when it comes to deciding which items are misfitting and should be abandoned or replaced, one should use the suggested cut-off scores as a guide, and then rely on his professional judgment and intuition to reach the best possible decision.

Smith (1996) provides a table of strings of responses to polytomous items together with the mean square fit values, the point measure correlation and a diagnostic comment for each string. The point measure correlations are similar to the point biserial correlations but correlate responses with Rasch measures instead of raw scores.

Standardized infit and outfit statistics

Wright and Masters (1982) suggest also standardizing these mean squares and transforming them into fit t-statistics by:

$$t_i = \left(u_i^{\frac{1}{3}} - 1\right) \frac{3}{q_i} - \frac{q}{3} \quad \text{or} \quad t_i = \left(v_i^{\frac{1}{3}} - 1\right) \frac{3}{q_i} - \frac{q_i}{3} \quad \text{where } q_i \text{ is the variance of the mean square.}$$

Karabatsos (2000) argued that the value of the t-statistic was sensitive to sample size and that reliance on this statistic could lead to the false detection of misfit. Also Li and Olejnik (1997), according to Curtis (2004), reported that all misfit indicators investigated (there were five misfit indicators) deviated substantially from a normal distribution raising questions about the transformation that is used for computing the t-statistics.

2.3.4.(iii) *Uses and criticisms of the infit and outfit mean square statistics*

Smith (2000) suggests that the infit and outfit person fit statistics can be used, just like the infit and outfit item statistics, in three different types.

First, the person total fit statistics, which is the sum of the chi-squares resulting from the encounter between any item and a given person.

Second, the person between fit statistic, which is based on some characteristic of the items that can be used to separate them into meaningful groups, like item difficulty, item type or cognitive level. This statistic has the potential to detect differences on performance over subsets of items.

Third, the person within fit statistic is used in conjunction with the person between fit statistic and is summed over all the items within a given item subgroup. This statistic allows for the identification of anomalous responses to a subset of items that might well be overwhelmed in the total fit statistic.

Smith (2000) however notes that most currently available Rasch calibration programs do not contain the person between fit statistic and have sacrificed an important tool in detecting measurement disturbances.

Infit and outfit statistics were designed to identify misfit with undifferentiated patterns of response and in the case of outfit, the presence of lucky guessing or carelessness. In addition to that, Wright (1997) suggests regressing residuals on item difficulty to bring out guessing or sleeping and on item position to identify fumbler or plodders and Hambleton et al. (1991) suggest standardized residuals against ability plots for assessing model-data fit.

Douglas (1990) comments on the common misapprehension that the standardized infit and outfit statistics have the power to detect all types of departure from the objective measurement model by writing

Not only should we not expect Z (standardized infit and outfit) to detect all aspects of misfit in persons, but any insistence on statistics that might claim such universality would be naïve. (p. 75)

He then answers to the criticism that Z does not detect a particular type of misfit by pointing out that the misfit investigations are usually induced artificially via specifically distributed simulated data, thus being “confirmatory” in contrast to the exploratory role for which Z was designed.

Douglas (1990) concludes that his research shows that at the exploratory level Z is quite satisfactory.

Many researchers have used simulated data, like Rudner (1983), Meijer et al. (1994).

The latter point out the following possible inefficiencies of such studies:

Although the theoretical framework is non-parametric the data are usually simulated using parametric IRT models. A standard normal distribution for the ability and a uniform distribution with equidistant item difficulties within a specified range, say $[-2, 2]$ are commonly used.

In practice this may easily not be the case.

Furthermore two assumptions are used about cheaters.

First these persons are assumed to answer the majority of the items in their own and only cheat in the very few hardest items. Second, cheating is assumed to always result in correct answers since it is done from more able persons. However in real situations desperate or anxious candidates may cheat from less able students and the more able students will not always answer the hardest items correctly even though they have a higher probability of doing so.

Finally, guessers are assumed to answer the items by randomly guessing the correct answer on each of the items with probability $\frac{1}{n}$, where n is the number of alternatives in a multiple choice test. However Hambleton (1993) notes that low-ability examinees score lower than they would actually score by randomly guessing. According to Hambleton (1993), Lord noted that this phenomenon could probably be attributed to the ingenuity of item writers who develop attractive but incorrect answer choices. On the other hand Nunnally and Bernstein (1994) distinguish two categories of guessing, the blind guessing, where guessers indeed guess randomly, and sophisticated guessing where the individual might not know which answer is correct but can improve his odds

by ruling out certain incorrect alternatives. Therefore in real situations it is not easy to distinguish how an individual guesses and opinions differ.

Further criticism of these statistics concerns their distribution and the fact that they are only approximately Chi-squares and whether the true distributional properties of these Chi-squares or their transformations were known (Karabatsos, 2000). Karabatsos argues that the distributional problem arises from the fact that the residual is the difference between an integer observed score and a non-integer expected score. He continues by saying that the use of the t-statistics for the infit and outfit mean squares is illogical. He then makes reference to Smith (1991) who showed that the distributions of the infit and outfit mean squares and the corresponding t-distribution are sensitive to sample size, test length and person ability and item difficulty distributions.

Curtis (2004, p. 130) argues that the method used by Karabatsos is flawed, because it does not simulate large samples of independent observations drawn from a population. This technique of repeating observations results in no change in the deviation from the mean but with an increase in N leads to reduced error variance and therefore artificially inflated t values. Curtis suggests that a better alternative would have been to identify the ability and difficulty distributions and to simulate data sets of increasing size based on those distributions and then to look at the trait distributions.

Nonetheless Curtis acknowledges that the t-statistics are sensitive to sample size and test length and possibly other variables and comments that this makes the use of the t-statistic in setting acceptance criteria for persons or items questionable.

Smith (1990) states that, since real data never fit any ideal model, all applications of Chi square are approximations and even though the mean square statistics are not true Chi squares they are regular enough to identify outliers reliably.

Another unresolved issue (Karabatsos, 2000) is the use of responses for both the parameter estimation and fit analysis. The responses are used to estimate item and person parameters. To calculate the residuals the expectation is needed which is a direct function of the parameters. Wright and Masters (1982) make reference to this point by stating that the estimated probability of success (P_{nik}) is used instead of the true probability (Π_{nik}), however doing so has proven quite satisfactory.

Other statistics that can be used efficiently with the Rasch model include the M-statistic for an examinee n , which is the sum of the product $X_{ni}\delta_i$ over all the i items (Molenaar and Hoijtink, 1990) and the I-statistic which measures the log-likelihood fit of an examinees responses with the predictions of an IRT model (introduced by Levine and Rubin).

Smith (1990) however concludes that the Wright-Panchapakesan approximations stand up well in comparison with possibly more precise tests such as likelihood-ratio Chi squares (Levine and Rubin, 1979) and the M-statistic (Molenaar and Hoijtink, 1990). Studies of the distributional properties of the Wright-Panchapakesan statistics show that the tails of their distributions are regular enough to identify outliers reliably. Therefore there is no practical reason to use anything more complicated.

Also, Meijer and Sijtsma (2001) comment on the fact that outfit and infit do not reflect the probability of ordering of the score patterns, by questioning whether this is relevant. They state "What is needed is an indication of how much misfit disturbs the estimated measures, not the likelihood of any particular score pattern" (Meijer and Sijtsma, 2001, p. 823).

Curtis (2004) concludes his literature review by saying:

Given the concerns raised by Karabatsos (2000) about the distributional properties of residual based fit statistics and about factors that influence them, there is a need to explore their distributions and the sample and item characteristics that might shape them in order to develop advice that is both soundly based and that is useful to practitioners. (p. 131).

2.3.5 Misfit as a threat to measurement

Many researchers (Athanasou & Lamprianou, 2002; Karabatsos, 2003; Reise & Flannery, 1996; Rudner, 1983) have argued that aberrant responses may lead to misleading score interpretations and consequently to invalid measurement.

Wright and Masters (1982) state:

If the fit statistics of a person's performance are acceptable, we say that their measure is "valid". (p. 114)

In discussing fit to the Rasch model, Smith (1990) raises two questions the first of which being about the overall fit of the data to the model. He then states:

The second question concerns the degree to which the total score that an examinee earns on a test adequately summarises the examinee's total set of responses. ... This is not a question of the utility of the data for analysis by the measurement model, but of the meaning (validity) of the measure for the individual. ... No matter how hard we try to construct potentially valid tests there will always be individual performances for whom the tests were not valid. (p. 78)

Smith (1986) also raises the question of whether an inconsistent individual (an individual with an aberrant response pattern) will exhibit such inconsistency in other testing situations.

Also, with regard to the influence of the infit and outfit mean square statistics, according to J. M. Linacre (personal communication, July 28, 2006):

Large outfit is a greater threat to the overall measurement system. Typical causes are careless mistakes and lucky guesses, but lucky guesses and careless mistakes are usually easy to diagnose, and to eliminate from the dataset, if desired.

Large infit is a greater threat to the validity of the individual person measures. Large infit can be caused by special knowledge and alternative

curricula. These are harder to diagnose. It is usually not clear how these affect pass-fail decisions and such like.

However, misfitting examinees are rarely a severe threat to overall measurement. If in doubt, analyze the dataset with and without them, and compare the item difficulties by cross-plot. It is unusual for there to be any distinguishable impact of the person misfit.

Linacre (2006) also emphasises the effect of high infit mean squares on items by explaining that these indicate that the items are mis-performing for the people on whom the items are targeted and this is a bigger threat to validity.

In a recent study, Lamprianou (2005) investigated whether the internal consistency (as measured by Cronbach's alpha) of the raw scores is smaller for groups of examinees with more misfitting response patterns. He also investigated whether the correlations of scores of examinees with misfitting response patterns have a lower correlation with other external measures of ability taken very close to the exam used for the measure of the ability (that is, whether there was a lower degree of concurrent validity).

He concluded that more misfitting response patterns lower the internal consistency of the raw scores, but no relationship was found between misfit and concurrent validity.

He then suggests that the absence of a relationship between misfit and concurrent validity could mean that either scores with aberrant response patterns do not lead to invalid interpretations, or because of a possible combination of aberrance in misfitting response patterns (for example, the raw score may be lowered by increased test anxiety and at the same time increased by special knowledge).

A further explanation could be that the same examinees consistently misfit, in the same way, in two successive tests measuring the same ability. (Lamprianou used tests from two different settings. One was the end of the year exam, taken by all students graduating from high school, and the other was the university entrance exams, consisting of two different tests. The tests were on the same syllabus, taken by more or less the same examinees and were only one or two weeks apart).

CHAPTER 3: METHODOLOGY

The data collection part of this study was spread over two academic years; therefore, the work was naturally divided into two phases.

Phase 1 involved administering a mathematics test and a test anxiety inventory to 572 students (age 15-16) in 5 schools in 3 different districts of Cyprus. An ADHD scale was also completed by the 13 teachers participating in the first phase, in which they had to rate the severity of ADHD symptoms of their students. This phase was planned mainly to investigate possible factors leading students to misfitting responses.

Finally, the internal consistencies of the raw scores, as measured by Cronbach's alpha, of fitting and misfitting students were compared with the use of confidence intervals for the alpha coefficient.

Phase 2 involved administering 2 mathematics tests, a mathematics self-concept questionnaire and a shorter version of the test anxiety inventory to 635 students in 3 different schools in two towns of Cyprus. The possible associations of math self-esteem and test anxiety with misfit were investigated.

Interviews of 21 of the most misfitting students were carried out in order to investigate further and in-depth the possible reasons for aberrant response patterns.

Furthermore, comparisons of proportions of fitting and misfitting students were made in order to investigate whether misfit is an inherent characteristic of students, that is, whether the same students misfit in administrations of different maths tests or in administrations of different psychometric scales.

The predictive validity of the scores of misfitting and fitting students in both maths tests were compared using correlation of their scores with other criteria.

Also, the internal consistencies of the raw scores, as measured by Cronbach's alpha, of fitting and misfitting students were compared with the use of confidence intervals for the alpha coefficient.

Following the comparisons of internal consistencies an investigation of infit and outfit was undertaken in order to assess the impact of unexpected responses to these mean square statistics.

3.1 Ethics

Before the commencement of the collection of data a letter was sent to the director of secondary education at the Ministry of Education and Culture (MOEC), asking permission to administer the mathematics test and the different scales to the students in the different lyceums. The letter also included the assurances of the researcher that the anonymity of the students and teachers involved would be safeguarded. Also, the researcher clarified that written consents for participation from the headmasters, teachers, students and their parents involved would be sought.

The director of secondary education gave the written permission (see appendix 1) for the realization of the study with the additional terms that

- no teaching time would be lost throughout the data collection and
- a final report with the results of the study would be sent to the MOEC and the Pedagogical Institute of Cyprus to enrich their library and to be used as a possible future reference.

Following the agreement of the researcher to adhere to all the terms, the researcher then asked for, and received, written consents for participation from the headmasters (see appendix 2) and teachers (see appendix 3) whose students would participate in the study. During this process detailed explanations were given to the headmasters and teachers involved both orally and in writing, about the purposes of the study and the role of the teachers in the process of data collection. The teachers then informed their students about this study, just before administering the test, and gave them a consent form to be completed by themselves and their parents.

In one school, this practice proved very time consuming and difficult (consent forms were lost in the process), therefore, in the remaining 4 schools only the written consent of the students was sought. At the same time, in order to accord to the assurances given by the researcher to the MOEC, all the students were asked to inform their parents about their participation in the study and if any parents objected the students could exercise their right to withdraw from the study (as it was clearly explained to them before giving their consent)

All the students willingly agreed to participate in the study and no objections from parents were brought forward.

The whole procedure followed a successful application to the ethics committee of Durham University for permission to proceed with the research.

Finally, permission for the use of the Test Anxiety Inventory (TAI, Spielberger, 1980) and items from the Self Description Questionnaire (Marsh and O'Neal 1984) was sought and granted from Mind Garden, the organization publishing the TAI, and Marsh respectively.

3.2 Phase 1

Three assessment instruments were used: a mathematics test, a test anxiety inventory (TAI) and an ADHD scale.

To overcome the problem of small numbers and unreliable results 25 classes in 5 different schools were selected giving a sample of 572 students. All students were attending the first form of the lyceum, ages 15 – 16.

Sampling (that is, the selection of schools, teachers and students) was based on the willingness of the 13 mathematics teachers who were involved to participate in the study.

The names given to the 5 schools, for the purposes of this study, were taken after the town which they belonged to. There were 3 schools in Limassol, named Limassol 1, Limassol 2, Limassol 3, and the other two schools were named Paphos and Dali, based on the towns in which they were located.

The researcher recruited four teachers, one from each of the four schools, all of which were at some point colleagues in the same school or friends of the researcher (the researcher was the fifth, being a maths teacher in one of the schools, Limassol 1). These four teachers, after being thoroughly informed by the researcher, orally and in writing about the purpose of the study, undertook to inform the other teachers in their school about the details of the study and to pass on the information material. All communication between the researcher and the schools was carried out through these four teachers.

3.2.1 The Maths test

The test (see appendix 1) was on straight line graphs, an algebra unit of the first form syllabus in the lyceums in Cyprus. It consisted of 12 multistep items carrying from 2 to 6 marks, giving a total score of 40. (The test is included in the appendices)

Crocker and Algina (1986) advise test developers to ask qualified colleagues to review the test items informally for accuracy, wording, grammar, ambiguities and other technical flaws. Following their advice, the researcher, who is an experienced teacher of mathematics and deliberately did not get involved in teaching first form students in the academic year 2004-2005, prepared the test with the help and suggestions for improvements from two other teachers working in two of the other schools involved. Once prepared, the test was then sent to all the teachers participating and their comments were sought. A couple of suggestions for the refinement of the test were brought forward, taken into consideration and the final refined test was prepared.

The test was administered over one 45-minute teaching period in January-February 2005.

It 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 researcher 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 researcher did not want to interfere with that.

The test was a typical classroom test for the following reasons:

- Its objective was to assess each student's ability on the specific unit and to identify possible weaknesses.
- It was prepared by mathematics teachers involved in the everyday teaching and was refined with the suggestions of other experienced colleagues.
- It was administered by the teachers, to their classes, during a normal 45-minute mathematics lesson.
- The class teachers marked it, returned it to their students and provided remedial instruction where they felt it was necessary.
- It was used as part of the assessment of students in mathematics for the second term of the academic year.

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.

3.2.2 Selection of the Rasch Models and fit statistics

Selection of the Rasch models

The Rasch models were selected from a large number of models offered by IRT for the analysis of the test data collected in this study for the following reasons.

- The Rasch models are the only IRT models that accept the raw scores of the examinees to be a sufficient statistic for the estimation of their underlying abilities thus maintaining the score order of students. Since raw scores are the basis for reporting results throughout the whole educational

system in Cyprus, and especially in classroom tests, this model is consistent with practice.

- The Rasch models are easier to work with, to understand and to interpret, because they involve fewer parameters.
- There are fewer parameter estimation problems than with the more general models.
- The Rasch models give stable item estimates with smaller samples than other IRT models.
- The person measures and item calibrations have a unique ordering on a common logit scale (Wright and Masters, 1982; Bond and Fox, 2001) making it easy to see relations between them. The item-person map provided by the Rasch software is very attractive to users.
- Validity and reliability issues can be addressed through the use of the Rasch models (this was discussed further in the literature review).
- The nature of the tests used in this study, the multistep mathematics problems, does not encourage guessing, therefore models that incorporate pseudo-guessing parameters are not appropriate for these data sets. The Rasch models assume no guessing.
- Finally, the wide use of the Rasch models and their fit statistics helps positioning this study within the literature and makes comparisons easier.

Selection of the fit statistics

Two fit statistics, the infit mean square (IMS) and the outfit mean square (OMS) have been used to estimate the degree of misfit of examinees in this study. These two fit statistics were preferred over a large number of fit statistics for several reasons:

First they have an exploratory nature (Douglas, 1990) and they can identify a wide range of potential sources of aberrance, like guessing, cheating, sleeping, fumbling, plodding and cultural bias (Wright, 1997). This exploratory nature is ideal for this kind of study where the identification of general aberrance is desirable. Furthermore, it is an advantage in the sense that a fit statistic that focuses on a specific type of aberrance may not have enough power to identify other types of misfit.

Second, the infit and outfit mean squares have been used successfully to assess the fit of the Rasch models for many years (e.g. Wright and Masters, 1982; Smith, 1990; Curtis, 2004), and this encourages their use in the context of the Rasch models.

Third, these statistics are computationally simpler and they stand up well in comparison with possibly more precise tests, therefore there is no practical reason to use anything more complicated (Smith, 1990).

Finally, they are utilized by most of the available software packages for Rasch calibrations (e.g. Quest, Winsteps, 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 contribute to or 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 that would be expected if the data and the model were perfectly compatible.” (Bond and Fox, 2001, p. 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 underfitting or 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. Bond and Fox (2001) suggest the same values as Wright et al. (1994) and Rudner, Skagg, Bracey and Getson suggest infit cut-off score of 1.2 for rejecting response strings manifesting more than 20% unmodeled noise (as reported in Wright, 1995. para. 7). Karabatsos (2000) also states that “Convention suggests that 1.3 defines the minimum critical value for OMS (outfit mean square) and IMS (infit mean square) for classifying a person or item as misfitting the model” (p. 155). Athanasou and

Lamprianou (2002) interpret person fit statistics larger than 1.3, in classroom assessment, as meaning that the pupil was probably mismeasured.

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.

The Type I error rate is the probability of falsely rejecting an item or person as not fitting the Rasch model. Smith, Schumacker and Bush (1998) (as reported in Smith, Rush, Fallowfield, Velikova and Sharpe, 2008) used simulated dichotomous data and found Type I error rates that were significantly lower than 0.05 for both infit and outfit using various ranges of critical values (0.7, 0.8, 0.9 – 1.1, 1.2, 1.3). Furthermore, the Type I error rates decreased for the outfit as sample size increased. Similarly, Karabatsos (2000) also used simulated dichotomous data with sample sizes of 150, 500 and 1000 and test lengths of 20 and 50. He showed that both infit and outfit are dependent on sample sizes but that for sample sizes above 150 the Type I error rates were below 0.05 for both mean square statistics for cut-off score of 1.2 or 1.3.

Whether simulation studies with a fixed Type I error are used, or the suggested reasonable cut-off values (which are rules of thumb) the decision as to which ones to use is arbitrary. Which ever method is used however, 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, p. 198).

For the purposes of this study, given the fact that:

- The researcher believes 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 identifying unexpected test takers' response patterns).

- Classroom (low stakes) tests were used

and following the suggestions of Wright et al. (1994), Bond and Fox (2001) and Karabatsos (2000) the conventional cut-off score of 1.3 for both infit and outfit statistics is used.

For the same reasons 1.5 is used as a cut-off score for the questionnaires used in this study.

Software used

All calibrations and test data analyses were conducted with the use of WINSTEPS (Linacre, 2005) and the statistical analyses and inferences with the use of SPSS.

3.2.3 Validity and reliability of the Maths test in phase 1

Cronbach's alpha was 0.91, much higher than the reliabilities of 0.60 to 0.80 suggested by Athanasou and Lamprianou (2002) for classroom tests. This is an indication of high internal consistency of the items that comprise the test.

Many different sources of evidence were collected to support the construct validity of the test:

First, factor analysis and second, principal components analysis of the standardized residuals (Linacre, 1998a) were performed in an attempt to investigate the structure of the data and to assess whether it is unidimensional.

Third, it is widely acceptable in the literature that to judge whether items adequately represent the performance domain (or the specific curriculum in the case of a classroom test) the judgments of a panel of experts is required. Therefore, a short questionnaire (see appendices 12 and 13) was administered to 6 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 with 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 included any items on skills not included in the syllabus.

Fourth, the results of the test were compared with the final exam results of the students, separately for each of four schools that participated in the study, since each school used its own final examination.

Finally, two comparisons of the item estimates from two different calibrations (using two different samples: first with different orders of the items in the tests, and second with different genders) were made in order to assess whether invariance holds. This would imply that the construct measured by the instrument has the same meaning to the groups.

Misfitting students

Misfitting students were identified using the above-mentioned cut-off scores for the infit and outfit statistics. The numbers and proportions of misfitting students were presented, together with comparisons of equivalent proportions from a simulation study.

3.2.4 Test Anxiety Inventory (TAI)

The TAI is a self-reporting psychometric scale, which was developed by Spielberger (1980) to measure individual differences in test anxiety as a situation specific personality trait. It consists of 20 items, asking respondents to describe how they generally feel. The items are answered using a 4-point Likert-style scale, scored from 1 to 4 (where 1 = almost never and 4 = almost always).

Three scores can be derived: Worry (8 items), Emotionality (8 items) and Total (all items combined). Worry is defined as “cognitive concerns about the consequences of failure”, Emotionality as “reactions of the autonomic nervous system that are evoked by evaluative stress” and Total as a composite of responses to all 20 items (Spielberger, 1980, p.1)

The TAI was translated into Greek (see appendix 7) by the researcher, with the help of a psychologist colleague, and the Greek version was translated back into English (see appendix 8) by an independent experienced teacher of English literature, who had not previously seen the English version of the TAI.

The two English versions of the inventory (the original and the one translated from the Greek version) were then compared making sure that the translation into Greek did not distort the content of the items.

Validity and reliability of TAI

The reliability and validity of the TAI scores is supported by several types of evidence provided in the test manual. The evidence published by the developers in the manual includes:

- Test-retest correlations of the Total score of 0.80 or higher over two week time intervals and 0.62 over a six month time interval. This was an indication of a high degree of reliability, which is important for a high degree of validity.
- Alpha reliability estimates of the Worry and Emotionality factors with median values of 0.88 and 0.90 respectively (for the various groups used in the original study of the TAI), indicating satisfactory internal consistency for the 8-item subscales. Alphas for male and female high school students were 0.86 and 0.89 respectively for the worry subscale and 0.90 and 0.91 for the emotionality subscale.
- Logical patterns of relation between TAI scores and other criterion measures, including positive correlations with six other measures of anxiety and low-to-moderate negative correlations with measures of study skills, intelligence and ability.
- Factor analysis of the 20 TAI items identifying the two strong, distinct factors of Worry and Emotionality.

In this study, alpha reliabilities were calculated and compared with the ones provided in the test manual. The correlation between the Worry and Emotionality subscales was also computed

Also factor analysis was used in an attempt to identify the same patterns, i.e. whether two factors are extracted with the 8 items loading significantly on the one factor and the other 8 items loading significantly on the second as suggested in the manual.

Furthermore, descriptive statistics from the TAI analyses were compared with the published analyses.

A short questionnaire (4 items) was attached to the TAI (as the final and separate section) to help the researcher collect information about students' grades in Greek language, the amount of time students spent studying for their mathematics homework, whether they take private tuition in mathematics and whether mathematics is one of their favourite subjects in school.

Misfitting students

Misfitting students were identified using appropriate cut-off scores for the infit and outfit statistics (infit/outfit > 1.5). The numbers and proportions of misfitting students were presented. Finally, a chi-square (contingency tables) test was performed to investigate possible association between misfit in the maths test and misfit in the TAI.

3.2.5 Assessment of Attention Deficit Hyperactivity Disorder (ADHD scale)

Towards the end of the academic year the mathematics teachers were asked to rate the severity of ADHD symptoms of their students using an 18-item rating scale that was based on the diagnostic criteria of ADHD (American Psychiatric Association, 1994) contained in the Diagnostic and Statistical Manual of Mental Disorders Version 4 (DSM IV).

This instrument was a scale based on dichotomous items on which teachers were asked to consider a series of criterion met if the behaviour had persisted for at least

six months and it was considerably more frequent than that of most other students of the same developmental level.

It is recommended that for students to be diagnosed as having ADHD they must meet at least 6 out of the 9 criteria relating to inattention for the Predominantly Inattentive subtype, and at least 6 out of the 9 criteria relating to hyperactivity and impulsivity for the Predominantly Hyperactive/Impulsive subtype. For the Combined subtype they must meet both of the above conditions.

The ADHD scale was translated into Greek (see appendix 11) by the researcher and back into English by an independent experienced teacher of English literature, who had not previously seen the English version.

The two English versions of the scale (the original and the one translated from the Greek version) were then compared making sure that the translation into Greek did not distort the content of the items.

In 4 classes (90 students) the ADHD scale was given also to the language teachers to assess the behaviour of their students. The numbers of criteria met by students, as assessed by the language teachers, were correlated with the ones from the mathematics teachers' assessments.

3.2.6 The investigation of factors associated with misfit

Students' abilities were divided into three groups: the low ability, the medium ability and the top ability for mathematics. This was done using 3 different sets of cut-off ability estimates: the 30th and 70th percentiles, the 20th and 80th percentiles and the 10th and 90th percentiles.

The test anxiety estimates of students were divided again into three groups using again the 3 different sets of cut-off scores as in the ability ranges. Low anxiety, medium anxiety and top anxiety groups were formed for each set of cut-off scores.

Apart from ability, test anxiety and ADHD symptoms (which were measured with the Test Anxiety inventory and the ADHD scale respectively), other factors were considered.

These other factors include:

- Different schools. Although the same curriculum is used throughout the schools in Cyprus, different schools can be considered as a factor since it appears in the literature as a possible source of misfit. However, any possible association between misfit and different schools can not, in this case, be attributed to different curricula.
- Different teachers. The different teachers involved, teaching the syllabus and administering and marking the tests, could be a factor relating to misfit. However, since the numbers of students corresponding to each teacher are small, one should be cautious in the interpretation of the results.
- Student and teacher gender.
- Language competency. The first term grade in Greek language of each student is used as a measure of language competency.
- Interest in mathematics. The maths teachers were asked to assess the interest their students showed in the subject, using a 3-point Likert scale where 1 = none, 2 = sometimes interested and 3 = always interested.
- Private tuition in mathematics. Students had to complete a very short questionnaire attached to the TAI asking them, among other things, whether they were taking private tuition in mathematics.
- Ability. The students have been grouped into high, medium and low scorers, depending on their ability estimates from the Rasch model calibrations.
- Atypical schooling. The number of unauthorized absences during the first term of the academic year was used as an indication of atypical schooling. One unauthorized absence in the schools in Cyprus stands for an absence from a 45-minute teaching period without any written justification, either from a parent or from a doctor. If a student has completed 42 – 50 unauthorized absences during the year he/she is not allowed to take the final exams in June and has to take them in September, whereas with more than 50 such absences he/she has to repeat the year.
- Item order. Although all the tests had the same items, those were given in two different orders, A and B. The 12 items of test A were laid out in 4 pages. In B the items in each of the 4 pages were exactly the same as the items in A but in reverse order. The researcher did not want to use a hard-

to-easy order for B because it is not common practice for classroom tests. The two different item orders were used for two reasons. First, to investigate whether different item orders affect misfit in the tests. Second, to minimize possible copying during the test since all students in a classroom sit in pairs. The mean scores of students on the two item orders were compared (20.33 for A and 19.88 for B) and no significant differences in the performances were found ($p = 0.65$)

- Study time. Students were asked to state in the short questionnaire attached to the TAI, how much time, in minutes, they usually spend studying mathematics every day.

Log-linear analysis was used to investigate possible association of these factors with misfit. (For details of the method see appendix 14)

Is misfit an inherent characteristic of students?

A Chi square test was performed comparing the proportions of fitting and misfitting students in the two instruments (maths test and TAI) administered to the students in this phase.

Internal consistency of raw scores of fitting and misfitting students

The internal consistencies of the raw scores, as measured by Cronbach's alpha, of fitting and misfitting students were compared. For the purpose of these comparisons the standard error of alpha and the confidence intervals were calculated using the method suggested by Iacobucci and Duhachek (2003).

3.3 Phase 2

Four assessment instruments were used: two mathematics tests, a mathematics self-esteem scale and a shorter version of the test anxiety inventory. The Rasch models were used for the analyses of the students' responses to all the instruments used in this phase.

For the validation of these 4 instruments various studies, for collecting validity evidence, were used. Three of the validation studies were used in all 4 instruments. These included:

- Principal components analysis of the standardized residuals, after the Rasch calibrations, as proposed by Linacre (1998a).
- A plot of the factor loadings (on the first dimension extracted, other than the dimension measured by the test) against item measures.
- Correlations of the instrument results with other criteria.

To avoid repetition this set of validation studies will be referred to as the **Standard Validation Studies**.

To overcome the problem of small numbers and unreliable results 25 classes in 3 different schools were selected giving a sample of 635 students. Sampling was based on the willingness of the 13 mathematics teachers who were involved to participate in the study. Most of the teachers involved in this second phase were the same as the ones in phase 1.

The schools used in this phase are 3 of the 5 used in phase 1; therefore the names given to the 3 schools were Limassol 1, Limassol 2 and Paphos, exactly as in phase 1.

3.3.1 The first maths test (*The Diagnostic test*) in phase 2

The first test was a 'diagnostic' test (see appendix 5), administered towards the end of September, the first month of the academic year 2005 – 06. Such a test is always administered at the beginning of the year in lyceums in Cyprus to all first form students, the newcomers to the schools and its purpose is to identify mainly the weaker students, the ones with difficulties in the very basics in mathematics. For this

reason it contains items on the basic skills and abilities, the ones teachers feel are the most important for students to possess in order to be able to follow the syllabus of the first form in the lyceum. Once the weaker students are identified, they are encouraged to take extra lessons in the subject. These lessons take place after school hours and are offered free of charge by the school.

This specific test was on the previous year's syllabus, on mathematical concepts considered basic for the new year's course. It consisted of 27 items carrying from 1 to 5 marks, giving a total score of 50. Three out of these items, items 2a, 2b and 2c, were multiple choice questions with three options to choose from, carrying one mark each.

The researcher 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 again 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.

The test was administered over one 45-minute teaching period in the last week of September 2005. 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 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.

A similar diagnostic test was also administered at the beginning of the year in Language. The researcher collected the answer sheets to these language tests in the Limassol 1 school and kept them for later use and in particular for using them together with the maths diagnostic test in a study of methods for detecting multidimensionality. Therefore, the researcher had the answers, at the item level, of 298 students on 55 items (27 from the maths and 28 from the language tests). These data were used for investigating whether PCA of the Rasch standardised residuals was more effective in detecting multidimensionality than PCA of the raw scores.

Reliability and validity of the first mathematics test in Phase 2

For the study of the reliability of the test two equivalent indices were used: the student reliability (this index is given by the Rasch analyses) and Cronbach's alpha.

For the validation of the test, the **Standard Validation Studies** have been used. The other criteria used for correlation with the test scores included the final maths exam and it 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 the gender of students) were made to check that invariance holds, implying that the construct measured by the instrument has the same meaning to the two groups.

3.3.2 The second maths test in phase 2

The second test (see appendix 6) used in this phase was another typical classroom test, 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.

The test was prepared and administered exactly the same way as the other two tests used in this project, with the cooperation of the researcher with teachers from the schools involved.

It was administered to 18 out of the 25 classes, that is 445 out of the 635 students who originally took the 'diagnostic' test. The reason for this smaller sample was that some teachers were not very willing to help the researcher further by administering this second test.

The test was administered over one 45-minute teaching period in February and March 2006. It 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 for the same reasons mentioned for the maths test in phase 1.

The test was again a typical classroom test for the reasons also explained earlier for the maths test used in phase 1.

To ensure more reliable results a detailed marking scheme was again prepared and thoroughly explained to and discussed with all the teachers so as to leave no questions or ambiguities.

Reliability and validity of test 2 in phase 2

For the study of the reliability of the test two equivalent indices were used: the student reliability (this index is given by the Rasch analyses) and Cronbach's alpha.

For the validation study of the test, the **Standard Validation Studies** have been used again. The other criteria used for correlation with the test scores included the final maths exam and it was again done separately for each school.

A content validity questionnaire was also used.

Furthermore comparisons of the item estimates from two different calibrations (using two different samples, based again on students' gender) to check whether invariance holds, implying that the construct measured by the instrument has the same meaning to the groups.

Finally, comparisons of ability estimates from the two maths tests used in this phase of the study were made strengthening even further the belief that the two tests indeed measure the same ability, mathematical ability.

Is misfit an inherent characteristic of students?

Chi square tests were performed comparing the proportions of fitting and misfitting students in:

- The two maths tests (the diagnostic and the second test).
- The two psychometric scales (the TAI and the MSES).

in an attempt to investigate whether the same students consistently misfit over administrations of maths tests or of psychometric scales.

3.3.3 The maths self-esteem scale (MSES)

The original Self Description Questionnaire (SDQ), according to Marsh and O'Neal (1984), was specifically designed to measure 3 areas of academic self-concept (reading, math, general school) and 4 areas of non-academic self-concept (physical abilities, physical appearance, peer relations, parental relations).

The original SDQ provided a basis for the design of SDQ III, which contained the 7 scales (except that the peer scale was divided into same sex and opposite sex scales) and additional scales for emotional stability, problem solving/creative thinking, general self, religion/spirituality and honesty/reliability.

Marsh and O'Neal (1984) demonstrated that responses to the SDQ III measure a consistent, distinct, and theoretically defensible set of 13 self-concept dimensions. The construct validity of the instrument was supported by the demonstration of logical patterns of relationships with relevant external criteria, which were significantly correlated with the areas of self-concept to which they are most logically related, and less correlated with other areas.

For the purposes of this study, 6 items (out of the original 10 in SDQ III) from the maths self-concept scale were chosen, the ones that could be more easily translated into Greek without losing meaning and the ones that the researcher thought would be more applicable in the Greek school environment (Permission from Marsh was obtained for using these items from his SDQ III).

The six-item MSES (see appendices 9 and 10) was administered to the students in the 3 schools by their teachers, during a normal math period and took about 5 minutes to

complete. The purpose of the scale was explained by the teachers and very few students opted out of answering it.

Given the fact that the scales were not completed anonymously, and in an attempt to ensure honest completion of them by the students, the researcher gave the teachers who administered the questionnaires the following instructions:

1. Ask the students to complete the questionnaire honestly.
2. Assure them that nobody other than the researcher will see the completed questionnaires.
3. Let all students place their completed questionnaire in one envelop, which after all are collected will be sealed in front of the students.

All the teachers, as far as they assured the researcher, followed the instructions to the letter.

Reliability and validity of the MSES

For the study of the reliability of the scale the student reliability (this index is given by the Rasch analyses) and Cronbach's alpha were used. Furthermore, the item total correlations were calculated and used as another indication of the degree of internal consistency of the test.

For the validation study of the scale, the **Standard Validation Studies** have been used.

The other criteria used for correlation with the MSES scores were measures of academic achievement. These measures included the diagnostic test, the second maths test, the maths final exam and the language final exam.

Principal components analysis of the raw scores was also carried out. This was done only because the researcher thought that since the original SDQ was analysed this way, to establish its validity, it would be a good idea to verify the unidimensionality of the scale using another well established method too.

Finally male-female comparisons were made revealing no differences in the MSES scores.

3.3.4 Shorter version of the Test Anxiety Inventory

To assess the test anxiety of the students a shorter version of the TAI (Spielberger, 1980) was used.

The researcher, in this phase, was not interested in breaking up test anxiety into the two factors but in measuring the students' test anxiety with a shorter, and easier to administer, questionnaire in an attempt to investigate whether test anxiety affects misfit in tests with multiple choice items.

The original TAI, which was used in phase 1 of this study, consisted of 20 items, asking respondents to describe how they generally feel. The items were answered using a 4-point Likert-style scale, scored from 1 to 4 (where 1 = almost never and 4 = almost always).

The shorter version of TAI was developed from the analyses of the original one administered in phase 1 and consisted of 10 items, aiming to measure the overall test anxiety of the respondents.

Out of the 8 items measuring the worry factor in the original TAI, 4 (the items with the highest loadings on the worry factor) were selected. Similarly, out of the 8 items measuring the emotionality factor, 4 were selected, again the ones with the highest loadings on the emotionality factor.

Finally, from the 4 remaining items on the original scale, which measure general anxiety, 2 were selected based on their infit and outfit values. The two items with mean square statistics closer to 1, the expected value of these statistics according to the Rasch model, were selected.

The researcher, in an attempt to achieve honest completion of the questionnaires by the students, gave the same instructions to the teachers who administered the TAI as the ones for the MSES. The researcher once again received the assurances of the teachers that instructions were followed to the letter.

Reliability and Validity of the TAI

For the validation study of the TAI, the following evidence was collected:

First, comparisons of short TAI results with the original TAI results from the first phase were made to see if similar results were obtained, especially the differences between male and female levels of anxiety.

Second, principal components analysis of the raw scores was performed.

Third, the **Standard Validation Studies** have been used once more and for correlations of the short TAI scores with other criteria, the maths test scores were used in an attempt to verify the significant negative correlation between test anxiety and test performance.

3.3.5 Predictive validity and internal consistency of scores of fitting and misfitting students

Predictive validity

The predictive validity of the scores of misfitting and fitting students in both maths tests were compared using correlation of their scores with other criteria. The other criteria used were the students' first term grade in maths, their maths final exam score and for the first maths test the scores on the second test and vice versa.

To make reliable comparisons 95% confidence intervals of the correlation coefficients were calculated using Fischer's transformation (which is explained in the results).

Internal consistency

The internal consistencies of the raw scores, as measured by Cronbach's alpha, of fitting and misfitting students in both tests were compared. For the purpose of these comparisons the standard error of alpha and the confidence intervals were calculated using the method suggested by Iacobucci and Duhachek (2003).

3.3.6 The Interviews

Following the calibrations of the second test in phase 2, misfitting students were identified with the use of the infit and outfit mean square statistics. From those students, 21 were selected to be interviewed. Those were the 21 students (out of the 34 most misfitting students) from the researcher's school. This number (21) represents approximately the 61.8% of the 34 students with the most unexpected responses, and this percentage is equivalent to the percentage of students in the sample that come from the researcher's school (59.8%, 266 out of the 445 students).

The researcher believed it was easier to interview the students during the morning, when all were in school; the reason for selecting misfitting students from the one school only was the easy access to the students and the ease with which the researcher could get the consent of the headmaster to interview the students and the consent of the teachers to allow students to leave their classes for a few minutes.

The interviews were planned to be semi-structured. The researcher set up a general structure by deciding in advance what ground was to be covered and what main questions were to be asked.

Then, the interview schedule was prepared having in mind the research questions, that is, the reasons for unexpected responses in classroom maths tests.

Part A of the schedule contained some general questions about the feeling of students about maths, whether they had confidence in the subject, whether they often make careless mistakes and the purpose was to make the interviewees feel more comfortable with the interview setting.

Part B contained the main questions, first about the test in general (i.e. whether it was easy or difficult, whether they had time to finish it and time to double-check their answers and whether they felt that there were any questions which in their opinion were not covered in the syllabus). Then each student was going to be asked about the question or questions on which his/her response was unexpected, with the aim to find the reasons behind this.

The students interviewed answered each question openly, sometimes at some length, in their own words and the interviewer responded with follow up questions to get the students to clarify or expand on the answers if necessary.

To ensure complete concentration and no disturbances by any teachers or students, the interviews took place in a small room, the office of one of the assistant headmasters who kindly agreed to offer it for the purposes of this study. Students were sent by their respective teachers to the 'interview room' during their lesson.

Given the confidentiality of the results of the analyses and of the identification of misfitting students all of the interviews were conducted during lessons other than mathematics so that the mathematics teachers of the interviewed students would have no way of knowing which of their students were identified as misfitting. Also, the selected students were allowed to leave their classes only if their teachers felt that in doing so, the loss of the 10-15 minutes from the lesson would not affect their performance.

Before commencing the interviews the researcher presented himself and explained thoroughly and in layman's language:

- The purpose of the interviews, being the in-depth investigation of unexpected response patterns
- How unexpected response patterns were identified
- Why these specific students were selected
- The confidentiality and anonymity of the process, giving assurances to the students that these interviews would in no way affect their school performance or school grade. Furthermore, they were reassured that their mathematics teachers had no knowledge of which students were selected and would certainly have no access to the interview material.
- The reason why the interviews had to be tape recorded.
- The choice they had to withdraw from the study, whenever they felt like it, without giving any reasons or having to suffer any consequences from the withdrawal.

After these explanations, the students were asked to sign a form expressing their consent to participate in the interviews, if they agreed and to be tape recorded.

All 21 students were very willing to participate and the interviews were conducted in a very friendly environment. The students answered all the questions, as far as the researcher could tell, honestly.

The first three interviews were used as pilot interviews, with which the researcher made sure that:

- Students were comfortable with his approach and explanations about the purposes of the study.
- The questions asked were clear.
- The tape recording worked properly, producing tapes that were easy for the researcher to transcribe from later and
- No disturbance was caused to the school and the learning process of the students involved in the interviews.

In order to make the material collected from the interviews manageable, the researcher transcribed them verbatim. Although some information, like body language or facial expressions is lost, the transcript provides a “true record of the original interview” (Derver, 1997).

The transcripts were written in Greek. The researcher then made, from the transcripts, a short profile for each student based on his/her answers to the general questions followed by a transcript of all the answers regarding possible reasons for the unexpected responses to some specific questions. The short profile and the shorter transcript were written in English this time, with a direct translation, by the researcher, from the original transcript.

To assist in the formulation of conclusions, the researcher presented the reasons for the unexpected responses in a tabular form. This table is presented in the section of the results.

Infit and outfit investigation

Following the comparison of internal consistencies and driven by the curiosity to explain why the internal consistencies were lower only for high infit values the researcher carried out an investigations into the effect of unexpected responses on the two mean square statistics.

First, the effect of one unexpected response on the outfit was considered at various test lengths and second the number of unexpected responses (which the researcher calls 'less likely' responses) needed to make the infit exceed the cut-off values thus characterising the response pattern as aberrant.

CHAPTER 4: RESULTS

The data collection part of this study was spread over two academic years; therefore, the analyses were naturally divided into two phases.

In phase 1 the maths test and the Test Anxiety Inventory (TAI) were calibrated using the Rasch models and misfitting students in the mathematics test were identified. Hence the proportions of misfitting students in each category of each factor under investigation were calculated and tests were carried out in order to infer whether there was association between the factor and misfit in the test. Furthermore the misfitting students in both the instruments were compared to see whether the same students misfit consistently. Finally confidence intervals for Cronbach's alpha were calculated in order to assess whether the internal consistency (as measured by Cronbach's alpha) of the raw scores is smaller for groups of examinees with more misfitting response patterns.

In phase 2 two maths tests, a short maths self-esteem scale and a shorter version of the Test Anxiety inventory were calibrated again using the Rasch models and students with aberrant responses in the tests were identified. Hence the consistency of misfit in mathematics tests and in psychometric scales, was investigated. Furthermore, possible associations between maths self-esteem or test anxiety and misfit in mathematics tests were investigated. Also correlation coefficients between test scores and other criteria were compared in order to assess whether the predictive validity of the score interpretations of misfitting students was lower than that of fitting students. Finally interviews of 21 students with unexpected responses were taken for an in-depth exploration of the reasons for misfit.

4.1 Phase 1 results

The sample

The maths test was administered to 572 students in 5 schools: Limassol 1, Limassol 2, Limassol 3, Paphos and Dali.

In Limassol 1, 3 teachers, 8 classes and 181 students were involved, in Limassol 2, 3 teachers, 6 classes and 136 students, in Paphos, 4 teachers, 5 classes and 123 students, in Dali, 2 teachers, 4 classes and 88 students and in the last school, Limassol 3, 1 teacher, 2 classes and 44 students. A total of 12 teachers and 25 classes were involved. The smallest number of students taught by a teacher was 23 (one class) and the largest was 68 (three classes).

Overall, out of the total of 572 students, 46.7% were male and 53.3% female.

The number of female students in the sample is greater than that of male students mainly because a much larger number of male students (than female students), after leaving the gymnasium, choose to attend a technical school rather than a lyceum.

Table 4.1.1 shows the distribution of the 572 students by gender, in the five different schools.

Table 4.1.1 Gender * school Crosstabulation

		School					Total
		Limassol 1	Limassol 2	Paphos	Dali	Limassol 3	
Gender	Male	93	61	55	35	23	267
	Female	88	75	68	53	21	305
Total		181	136	123	88	44	572

4.1.1 The Maths Test

Test calibrations

The Rasch PCM model was used for the calibrations. The first calibration on the full dataset revealed five misfitting items ($1.3 < \text{outfit} < 1.98$) and 16 badly misfitting students ($\text{outfit} > 3.0$).

The 16 students were removed and a second calibration was performed, revealing only 4 slightly misfitting items. Those items were retained in the dataset (the reasons for not removing the items are explained).

The item statistics from the second calibration were then used for the final calibration in order to obtain the students statistics.

Item-person maps are presented to show how well the items were targeted for the population of students and finally the students were divided into groups according to their ability for investigating later on whether ability is associated with misfit.

First calibration

The first calibration, in which the full set of the test data was used (12 items and 572 students), revealed two badly misfitting items, items 1 and 11 ($\text{outfit} > 1.5$) and 3 slightly misfitting items, items 9, 2 and 3, ($1.3 < \text{outfit} < 1.5$) as shown in table 4.1.2. Also two of those items had infit of 1.44 and 1.39. The mean values of infit and outfit were 1.02 and 1.11 respectively.

It is worth noticing that the most misfitting items were the ones with the lowest correlation coefficient with the total score (0.50 and 0.59, which are still significant) and the ones identified by factor analysis as having the smallest loadings on the dimension measured by the test. Also item 11 was the hardest item on the test (measure 1.19) and item 1 the second easiest (measure -0.76).

Table 4.1.2 ITEMS STATISTICS: MISFIT ORDER

ENTRY	RAW					INFIT	OUTFIT	PTMEA	
NUMBER	SCORE	COUNT	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	ZSTD	CORR.
11	370	532	1.19	.06	1.29	3.4	1.97	5.9	A .59
1	762	532	-.76	.07	1.44	6.1	1.59	3.5	B .50
9	575	532	.58	.05	1.39	4.9	1.45	2.4	C .63
2	1418	532	-.46	.04	1.25	3.3	1.35	2.6	D .69
3	1288	532	-.28	.04	1.17	2.4	1.33	2.3	E .69
5	626	532	-.25	.06	1.10	1.9	1.22	2.2	F .58
10	1268	532	.50	.04	.90	-1.6	.88	-1.6	f .80
4	1027	532	-.43	.05	.87	-2.2	.87	-1.0	e .71
8	1247	532	-.97	.06	.83	-2.2	.77	-1.5	d .69
12	768	532	.92	.04	.73	-3.9	.67	-4.1	c .78
7	780	532	.18	.05	.68	-6.3	.65	-5.1	b .77
6	611	532	-.20	.06	.61	-8.6	.55	-6.5	a .75
MEAN	895.	532.	.00	.05	1.02	-.2	1.11	-.1	
S.D.	327.	0.	.64	.01	.27	4.4	.42	3.7	

Table 4.1.3 shows the top part of the table with the student statistics in misfit order. This part of the table comes from the original calibration and shows students whose infit and/or outfit is greater than 1.8. The 16 most misfitting students (outfit and/or infit > 3.0) are shown in bold.

Table 4.1.3 STUDENT STATISTICS: MISFIT ORDER

ENTRY	RAW					INFIT	OUTFIT	PTMEA	
NUMBER	SCORE	COUNT	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	ZSTD	CORR.
365	3	12	-1.66	.50	2.44	1.5	9.90	4.4	A-.56
265	33	12	1.26	.34	2.59	2.4	7.09	3.3	B-.26
80	28	12	.76	.30	3.12	3.0	5.82	3.7	C-.14
256	36	12	1.68	.42	1.15	.4	5.48	2.4	D-.18
262	4	12	-1.45	.43	1.16	.5	5.01	2.6	E-.44
556	30	12	.95	.31	2.89	2.8	4.79	2.9	F-.18
194	30	12	.95	.31	2.41	2.3	4.63	2.9	G-.30
208	35	12	1.52	.38	1.55	1.0	4.62	2.3	H-.15
431	29	12	.85	.31	2.52	2.4	4.36	2.9	I-.15
485	8	12	-.91	.32	1.39	.9	4.25	3.0	J-.26
165	12	12	-.55	.28	1.34	1.0	4.07	3.6	K-.09
193	32	12	1.15	.33	1.12	.4	3.86	2.3	L-.42
271	32	12	1.15	.33	1.61	1.2	3.31	2.0	M-.02
259	8	12	-.91	.32	1.84	1.6	3.26	2.4	N-.12
217	15	12	-.32	.28	1.54	1.5	3.21	3.2	O-.11
257	8	12	-.91	.32	1.08	.3	3.16	2.3	P-.04
354	27	12	.66	.30	2.58	2.4	3.00	2.3	Q-.16
386	31	12	1.05	.32	2.04	1.9	2.99	1.9	R-.00
255	31	12	1.05	.32	1.78	1.5	2.97	1.9	S-.10
449	12	12	-.55	.28	1.75	1.9	2.94	2.7	T-.09
424	37	12	1.88	.48	1.30	.6	2.92	1.4	U-.09
196	8	12	-.91	.32	1.06	.3	2.91	2.2	V-.15
373	39	12	2.67	.88	.86	.4	2.86	1.3	W-.20
426	20	12	.06	.28	1.87	1.8	2.85	2.8	X-.02

Second calibration

All 16 students with outfit > 3.0 (2.8%) were considered badly misfitting and a threat to the measurement process and were removed, leading to a second calibration with again the 12 items, but this time with 556 students.

Table 4.1.4 shows the item statistics from this second calibration in misfit order (based on outfit).

Table 4.1.4 ITEMS STATISTICS: MISFIT ORDER

ENTRY	RAW					INFIT		OUTFIT	PTMEA	
NUMBER	SCORE	COUNT	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	ZSTD	CORR.	items
9	553	516	.59	.05	1.43	5.3	1.49	2.5	A .63	item 9
11	342	516	1.28	.06	1.24	2.7	1.44	3.0	B .62	item 11
1	749	516	-.83	.07	1.44	6.0	1.35	2.0	C .51	item 1
3	1251	516	-.30	.04	1.21	2.9	1.42	2.8	D .69	item 3
5	604	516	-.25	.06	1.13	2.2	1.29	2.7	E .58	item 5
2	1390	516	-.50	.04	1.21	2.7	1.07	.6	F .70	item 2
10	1212	516	.54	.04	.93	-1.1	.91	-1.2	f .81	item 10
4	997	516	-.45	.05	.90	-1.6	.92	-.5	e .70	item 4
8	1218	516	-1.02	.06	.83	-2.2	.63	-2.5	d .69	item 8
12	734	516	.96	.04	.74	-3.8	.68	-3.9	c .79	item 12
7	756	516	.18	.05	.69	-5.9	.66	-4.8	b .77	item 7
6	591	516	-.21	.06	.62	-8.1	.57	-6.0	a .74	item 6
MEAN	866.	516.	.00	.05	1.03	-.1	1.04	-.5		
S.D.	322.	0.	.68	.01	.27	4.2	.34	3.1		

This time there were only 4 slightly misfitting items. The mean values of infit and outfit were 1.03 and 1.04 respectively. The mean outfit value is much closer this time to the expected value of 1.

A summary of the results of the Rasch analysis from the second calibration is given in table 4.1.5

Table 4.1.5 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)
Examinees	556	0.12 (1.14)	-2.58 to 2.76	0.86	2.50	1.03 (0.46)	1.04 (0.64)
Items	12	0.0 (0.68)	-1.02 to 1.28	0.99	11.87	1.03 (0.27)	1.04 (0.34)

The range of student abilities was from -2.58 to 2.76, with a mean of 0.12 (SD = 1.14). The reliability of student estimates was 0.86. 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. The student separation index was 2.50. This indicates the spread of person measures in standard error units, in this case in 2.5 standard errors. The higher the value of the separation index, the more spread out the persons are on the variable being measured. A student separation index of 2.5 also indicates approximately 4 statistically distinct strata ($\text{strata} = 3.7$) of student abilities identified by the instrument ($\text{Strata} = [4(\text{sep. index}) + 1]/3$, Wright and Masters, 1982).

The item estimates ranged from -1.02 to 1.28 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 11.87, indicating that the spread of item estimates is about 12 standard errors.

Further investigation into the slight misfit of the 4 items showed that:

Item 1 was the second easiest and least discriminating question in the test. It was a very simple question asking students to just plot the point with coordinates (-2, 3) on a set of axes that was provided. It was so unexpectedly easy that many students from throughout the distribution of abilities managed to get it wrong, mainly because they added a line onto the diagram, while all they were expected to do was to plot a point.

Item 3 was just below average difficulty, and was on the most basic skill required in the chapter on straight lines; it asked students to 'Draw the line with equation $y = 2x - 3$ on the axes provided'. Given that this was the most typical and expected question in the test and the fact that more than half of the students (58%) take private tuition in maths (where they practice a lot the more 'standard' questions) most of the students did well, some even better than expected, thus making the item slightly misfitting ($\text{outfit} = 1.42$).

Item 9, one of the harder items in the test (measure 0.59), was asking students to 'Find the equation of the line which passes through the point (1, -2) and is parallel to the x-axis'. Although students were familiar with this type of question, what put some of the high scorers off was the fact that the line was parallel to the x-axis, with gradient zero, as opposed to the usual inclined line. Therefore a few of the high scorers missed that item causing it to misfit.

Finally, item 11 was the hardest item in the test (measure 1.28) and the most original and unexpected. Only the students with the highest ability and the ones who really understood the meaning of the 'gradient of a straight line' answered it correctly. This question had two parts. In the first part students had to choose the correct answer from 3 options (for 1 mark) and in the second they had to explain their choice (for two marks). The misfit in this item was most probably caused by the fact that although it was the hardest item in the test the first mark could be obtained by guessing and some of the lower ability students did indeed guess the answer.

Despite the slight misfit of these items, none of them were removed because the first two were considered to be basic and important for the test, and the other two, especially item 11, were very original items, which tested the ability of candidates to face novel situations.

Third and final calibration

The item statistics from the second calibration were then used for the third and final calibration which included the 12 anchored items and all the 572 students.

Figure 4.1.1 shows the item-student map. One can see that the test items are well targeted for students with abilities from 1 standard deviation below to 1 standard deviation above the overall mean ability. That is, the test items are well targeted for approximately the central 70% of the distribution of students' abilities. There are no items well targeted for the clusters of students at the very top (the high ability students) and very bottom (the low ability students) of the map.

Figure 4.1.1 Item – Student map

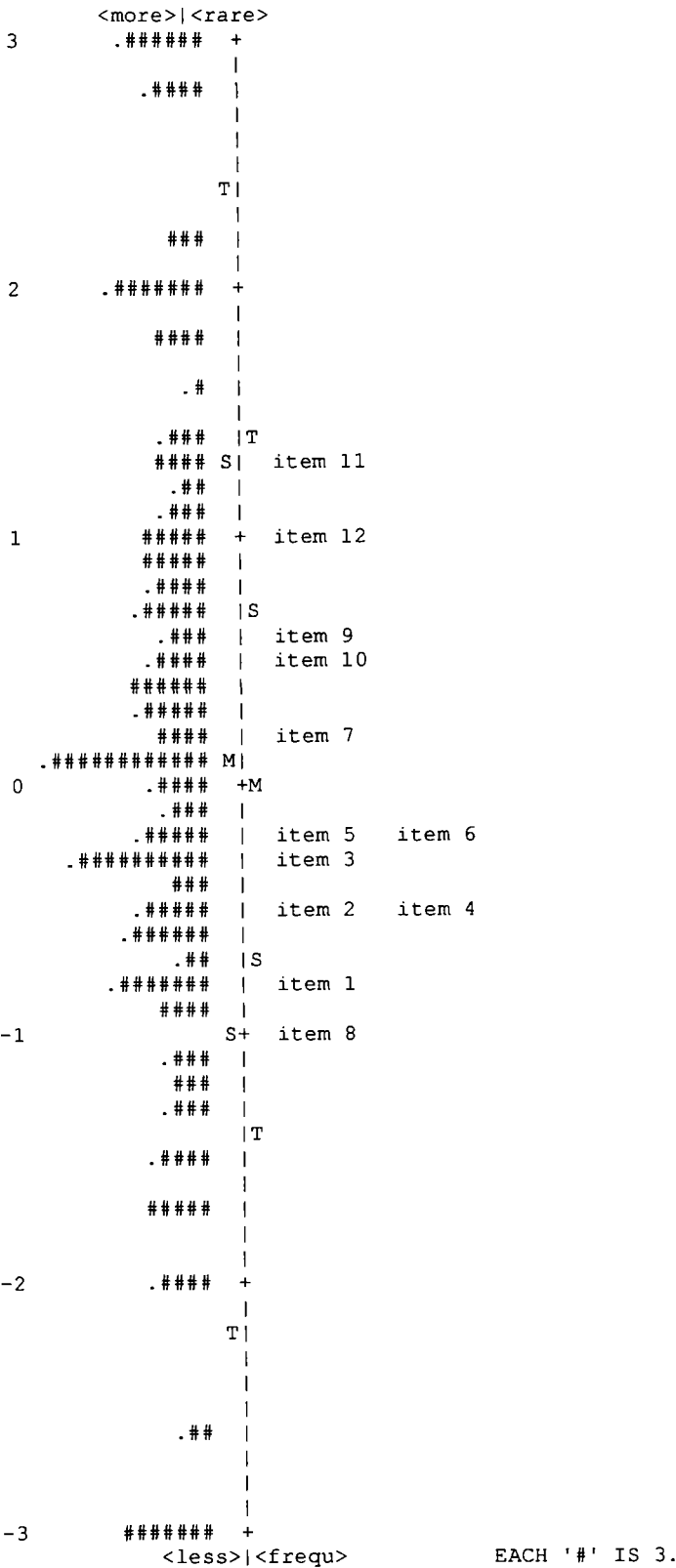


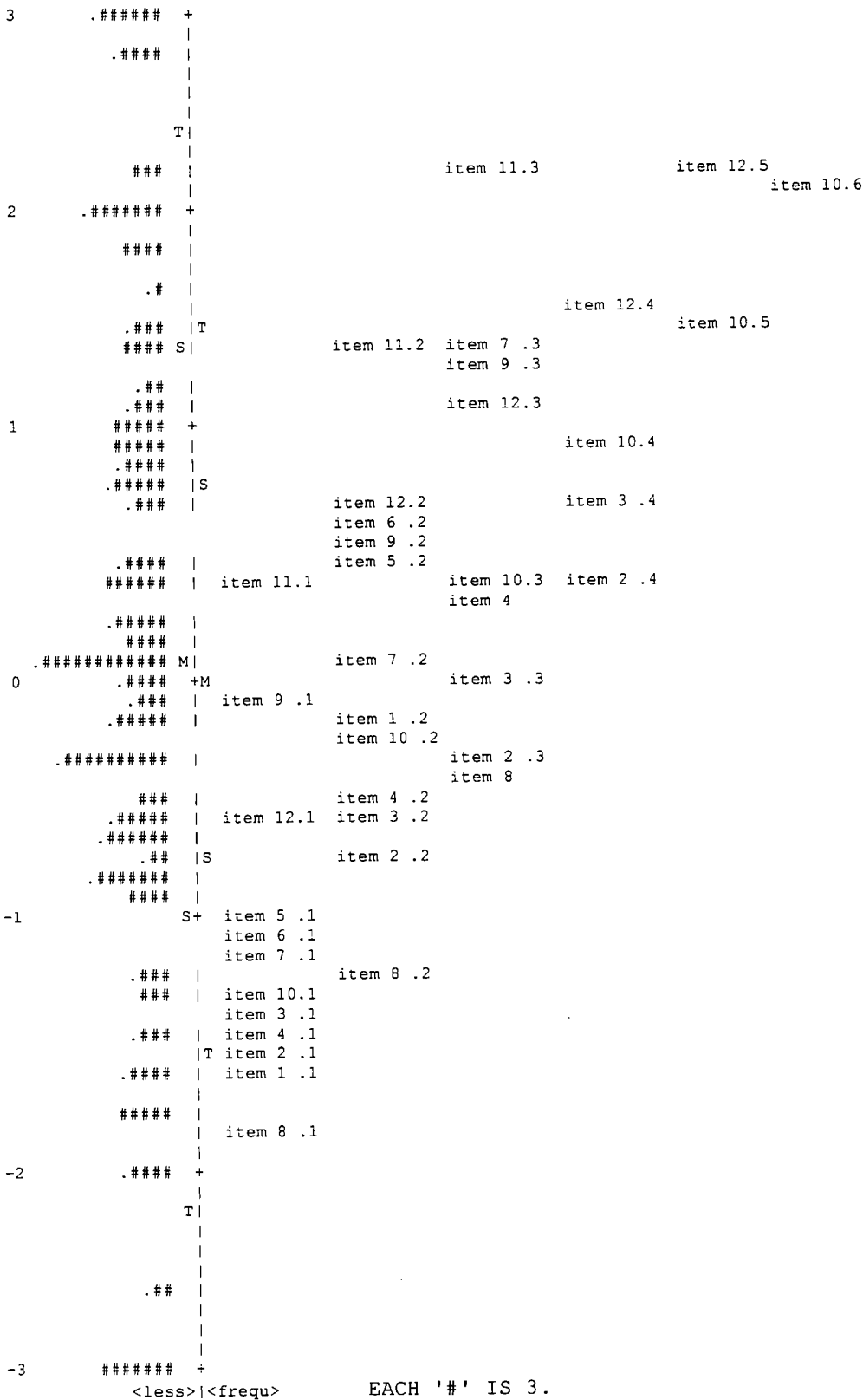
Figure 4.1.2 shows another item – student map, with the same items but this time with all the categories of the items (the thresholds for all the possible scores for each item).

It is obvious that the various steps of the items are well targeted for a wider range of abilities, from 2 standard deviations below to 2 standard deviations above the overall mean ability. Especially for students with low ability estimates, the first marks of items 8, 1, 2, 4, 3 and 10 could have been obtained. The bottom cluster of students was not very well targeted by the item steps but that cluster contains 21 students, which is a small proportion of the students in the sample (3.67%).

With a classroom test, which can not contain a large number of items because of the type of items used (multistep problems) and the limited duration of the test administration (45 minutes), the targeting of the items was satisfactory.

The two clusters at the top and the bottom of the figure represent the 19 students who scored full marks (40 marks) and the 21 students who scored no marks. These students are removed from the calibration process since their response pattern contains no information relative to the test items to estimate their ability (they are beyond the reach of this test). To provide a guide to possible ability estimates the logit estimate is based on a score of 1 for zero scores and a score of 39 (maximum possible score – 1) for the perfect score of 40. In this case a possible ability estimate for the perfect score is 3.74 (for 39 out of 40 it is 2.73) and for the zero score – 3.64 (for 1 out of 40 it is – 2.57).

Figure 4.1.2 Item – Student map (with item score thresholds)



Explanations into how the thresholds (boundaries between adjacent categories) are conceptualized in Rasch measurement are given in section 4.2.4, the phase 2 results, where the Rasch analyses of the MSES are presented. This particular instrument was chosen because, in the opinion of the researcher, it is easier to explain thresholds in the case of a short rating scale (MSES consists of 6 items) with the use of the RSM.

Different ability groups

For the purposes of further investigations, the range of abilities was divided into three different groups, the low, medium and top ability groups using three different cut-off scores (All the students were put into these three categories, even the top and low scorers, which although their ability was not accurately estimated there was no doubt as to which group they belonged).

First, the range of abilities was divided into 3 groups using the 30th (measure of -0.4984) and 70th (measure of 0.7804) percentiles. The lowest 30% of the distribution was labelled the 'Low 30% Ability' group, the middle 40% the 'Medium Ability' group and the top 30% the 'Top 30% Ability' group.

Second, the range of abilities was divided into 3 groups using the 20th (measure of -0.9353) and 80th (measure of 1.1905) percentiles. The lowest 20% of the distribution was labelled the 'Low 20% Ability' group, the middle 60% the 'Medium Ability' group and the top 20% the 'Top 20% Ability' group.

Third, the range of abilities was divided into 3 groups using the 10th (measure of -1.6949) and 90th (measure of 1.9318) percentiles. The lowest 10% of the distribution was labelled the 'Low 10% Ability' group, the middle 80% the 'Medium Ability' group and the top 10% the 'Top 10% Ability' group.

4.1.2 Reliability and validity of the test

For the study of the reliability of the test two equivalent indices were used: the student reliability (this index is given by the Rasch analyses) and Cronbach's alpha. Furthermore, the item-total correlations were calculated and used as another indication of the degree of internal consistency of the test.

For the validation study of the test, the following evidence was collected:

- *Analysis of a content validity questionnaire.*
- *For the investigation of the dimensionality of the test the following procedures were employed:*
 - o *Factor analysis, together with a scree plot.*
 - o *Principal components analysis of the standardized residuals after the Rasch calibrations, as proposed by Linacre (1998a).*
 - o *A plot of the factor loadings (on the first dimension extracted, other than the dimension measured by the test) against item measures.*
- *Correlations of the maths test scores with the final maths exam scores.*
- *Comparisons of the item estimates from two different calibrations (based on the order of the items in the test and on students' gender) to ascertain whether invariance holds.*

Reliability

The student reliability was 0.86. This index is an indication of the precision of the instrument and shows how well the instrument can distinguish individuals.

Cronbach's alpha was high (0.906) indicating also a high degree of reliability (such alpha is acceptable even for high stakes tests). Alpha is a measure of the internal consistency of the test.

Although both alpha and student reliability are estimates of the reliability, they differ slightly for two reasons. First, alpha is calculated using the raw scores which are not

linear measures and second, in the calculation of alpha the students who scored full marks or zero marks are included, whereas for the student reliability they are not.

Table 4.1.6 shows the item – total correlations.

Table 4.1.6 Item – total correlations

Items	Corrected Item – total Correlation
1	0.43
2	0.64
3	0.67
4	0.70
5	0.55
6	0.79
7	0.80
8	0.64
9	0.59
10	0.77
11	0.50
12	0.75

All items are good discriminators (correlations between 0.43 – 0.80), which is very satisfactory bearing in mind that the Rasch models require items with similar discriminations. Although items 1 and 11 are the ones with the lowest discriminating power, correlations of 0.43 and 0.50 are highly significant and considered satisfactory. These more traditional statistics were calculated in order to show the similarity between these and the Rasch statistics. Both methods have identified items 1 and 11 as the least discriminating.

Validity

A short questionnaire on content validity was administered to 6 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 included any items on skills not included in the syllabus.

Table 4.1.7 shows the number of experts who selected each option in each of the six statements.

Table 4.1.7 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	0	0	6
All the questions are clear and unambiguous	0	0	0	6
Students who know the answers have enough time to finish the test	0	0	4	2
All the important abilities and skills of the unit are assessed by the test	0	0	0	6
No irrelevant topics are included in the test	0	0	2	4
The test content is representative of the unit content as described in the curriculum	0	0	0	6

It is clear that all the experts agree or absolutely agree on all the statements regarding the content validity of the test.

Factor Analysis

Principal components analysis was performed using SPSS extracting only one factor. Table 4.1.8 shows the total variance explained by this factor, as well as the variance explained by all the other factors which are not significant.

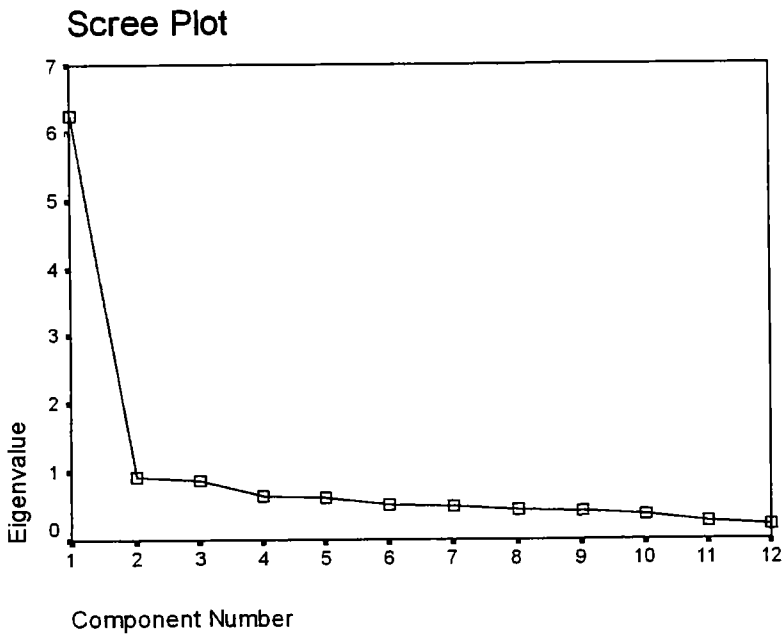
Figure 4.1.3 is the corresponding scree plot, the plot of the eigenvalues of the factors extracted.

Table 4.1.8

Total Variance Explained						
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6,250	52,079	52,079	6,250	52,079	52,079
2	,936	7,797	59,876			
3	,864	7,202	67,078			
4	,653	5,446	72,524			
5	,608	5,071	77,595			
6	,516	4,300	81,895			
7	,489	4,077	85,971			
8	,442	3,685	89,656			
9	,418	3,480	93,136			
10	,364	3,035	96,172			
11	,265	2,206	98,377			
12	,195	1,623	100,000			

Extraction Method: Principal Component Analysis.

Figure 4.1.3



The table suggests that the test measures only one ability, which accounts for 52% of the variation in the data. The scree plot also shows that only one factor has an eigenvalue greater than 1 and therefore the test can be considered unidimensional. The loadings of all the items on this factor are significant (from 0.482 to 0.851), strengthening further the belief of a unidimensional test.

Principal components analysis of the standardised residuals

Principal components analysis (PCA) on the standardised residuals (Linacre, 1988) was performed in WINSTEPS yielding:

PRINCIPAL COMPONENTS (STANDARDIZED RESIDUAL) FACTOR PLOT				
Factor 1 extracts 1.8 units out of 12 units of item residual variance noise.				
Yardstick (variance explained by measures)-to-This Factor ratio: 50.4:1				
Yardstick-to-Total Noise ratio (total variance of residuals): 7.7:1				
Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)				
		Empirical	Modeled	
Total variance in observations	=	104.2	100.0%	100.0%
Variance explained by measures	=	92.2	88.5%	88.9%
Unexplained variance (total)	=	12.0	11.5%	11.1%
Unexpl var explained by 1st factor	=	1.8	1.8%	

The variance explained by the measures (i.e. by the dimension measured by the test) is 88.5% of the total variance. It is also more than 50 times the variance explained by the first factor extracted by PCA on the standardised residuals and about 8 times the total unexplained variance in the data. The unexplained variance is 11.5% of the total variance in the data.

Also, the variance explained by this first factor is 15% of the unexplained variance (1.8 out of 12), but that is just 1.8% of the total variance in the data.

All of the above support the hypothesis that there is no second dimension present in the data, therefore the test is unidimensional.

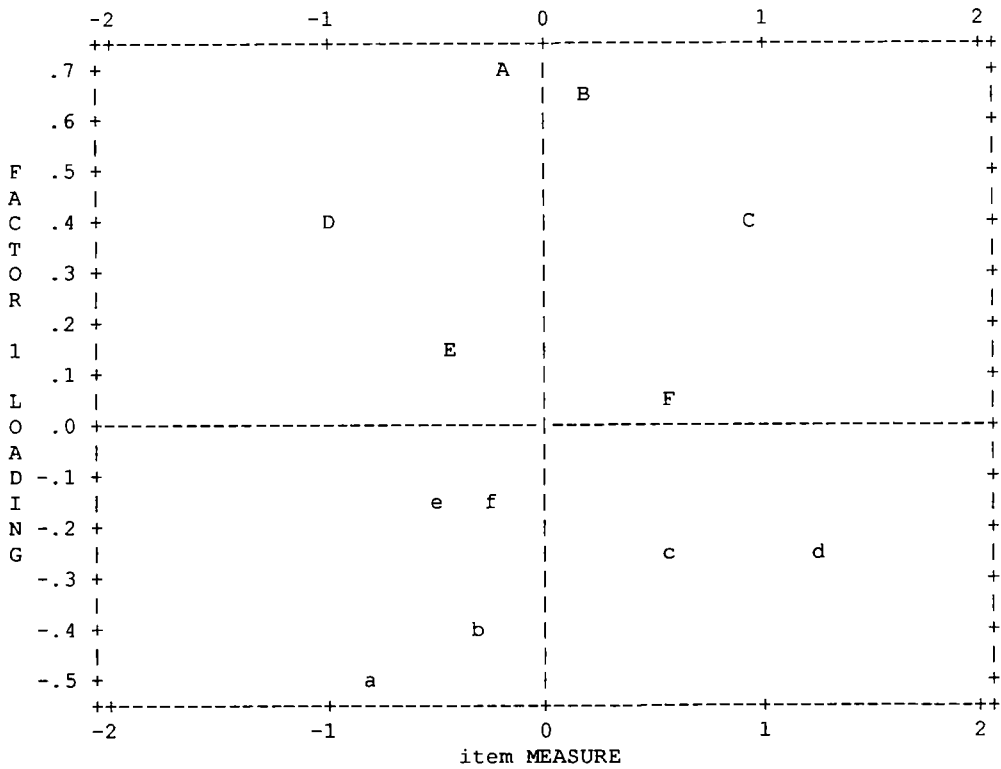
At first sight there seems to be some sort of discrepancy between the results of factor analysis of the observed scores and principal components analysis of the standardised residuals. The first method extracts a factor which 'explains' 52% of the variance in the data whereas in the second method the variance 'explained' by the measures is 88.5%. Factor analysis extracts factors based on the intercorrelations between the scores on the items. This method can be misleading when there are a few highly correlated factors which may be identified and treated as different dimensions.

Also different response styles, different content areas, different item formats could define different dimensions which in a factor analysis could be extracted as minor factors or add to the unexplained variance if they are not significant factors.

The Rasch model on the other hand, constructs a unidimensional measurement system regardless of the dimensionality of the data. Then, the residuals should represent random noise, which when standardised would follow a normal distribution. Furthermore, the residuals would be independent of each other. As a consequence all elements in inter-item residual correlation matrix would be zero if the data fit the model. However, each observation, will to some degree, contain its own characteristic features. Principal component analysis of these standardised residuals identifies characteristics shared in common among items. These are often indications of secondary structures or sub-dimensions within the data that may warrant action and diagnosis. Therefore, according to Schumacker and Linacre (1996) Rasch analysis excels at aiding the “identification of the core construct inside a fog of collinearity” (p. 470)

This belief in a unidimensional assessment is strengthened further by figure 4.1.4 below, which shows the plot of the items’ loadings on the first factor extracted against the items’ measures.

Figure 4.1.4 Factor loadings against item measures.



According to Linacre (1998) the presence of item groupings in this table may be evidence of a second dimension. However, one can safely say that there are no obvious item groupings in this plot

Correlations of test scores with exam scores

The scores on the test were compared with the final mathematics exam results of the students in 4 of the 5 schools. This was done separately for each school since each school prepared its own final examination. The correlation coefficients (all highly significant) were:

- Limassol 1: $r = 0.76$ ($N = 181$)
- Limassol 2: $r = 0.78$ ($N = 136$)
- Paphos: $r = 0.88$ ($N = 123$)
- Limassol 3: $r = 0.84$ ($N = 44$)

Comparisons of item estimates from two calibrations

(a) Split of the data by item order

Finally the full set of data was divided into two subsets. Those were labelled ‘subtest A’ (consisting of the responses of 290 students in the test with item order A) and ‘subtest B’ (consisting of the responses of 282 students in the test with item order B). Separate Rasch calibrations on the two subtest data were conducted and table 4.1.9 shows the results of these calibrations.

From the calibration of subtest A 25 students were removed (14 maximum scorers and 11 zero scorers) leaving the responses of 265 students. From the calibration of subtest B 15 students were removed (5 maximum scorers and 10 zero scorers) leaving the responses of 267 students.

The second and third columns of the table give the raw score on each item, which given the sample sizes of 265 and 267 can easily be compared.

The last two columns of the table give the item measure and in bracket the standard error of this measure.

Table 4.1.9 Raw scores and item measures from the two calibrations

Items in difficulty order based on A	Raw Scores		Item measure (standard error)	
	Subtest A	Subtest B	Subtest A	Subtest B
Item 11	199	171	1.07 (0.08)	1.32 (0.08)
Item 12	388	380	0.87 (0.06)	0.96 (0.06)
Item 9	279	296	0.60 (0.07)	0.55 (0.07)
Item 10	635	633	0.48 (0.05)	0.52 (0.05)
Item 7	391	389	0.16 (0.07)	0.21 (0.07)
Item 5	303	323	- 0.19 (0.09)	- 0.31 (0.09)
Item 3	642	646	- 0.21 (0.09)	- 0.26 (0.06)
Item 6	303	308	- 0.31 (0.06)	- 0.20 (0.09)
Item 4	505	522	- 0.42 (0.07)	- 0.45 (0.07)
Item 2	712	706	- 0.50 (0.06)	- 0.42 (0.06)
Item 1	362	400	- 0.63 (0.09)	- 0.91 (0.10)
Item 8	615	632	- 0.92 (0.08)	- 1.02 (0.08)

Two things are worth noticing from the table above.

First, the item measures from the two calibrations are almost identical (within standard error). Only two items seem to have differences slightly large, items 11 and 1. The most probable cause of this difference with these two items is the order in which they appeared in the two subtests. Item 11 was the one before the last in subtest A but it was the last in subtest B. Being the last in subtest B it was probably not attempted by more students than subtest A, who probably, just by looking at it thought it was too difficult to attempt. This was also probably the reason why subtest A had 14 maximum scorers whereas subtest B only 5.

Also more marks were scored by the students in subtest A (199) than in subtest B (171) in this 3-marks item. Item 1 on the other hand had the opposite effect. It was the first in subtest A and the fourth in subtest B. Some students in subtest A probably answered it

too rashly, and carelessly feeling that it was too easy thus the large difference in the raw scores for this 2-marks item (362 for subtest A and 400 for subtest B).

Second, the order of the items in the two calibrations is almost the same. The first five items, in difficulty order from most difficult to easiest, are exactly the same. Any differences after that (items 5 and 6 or items 4 and 2) are so slight that within standard errors are negligible.

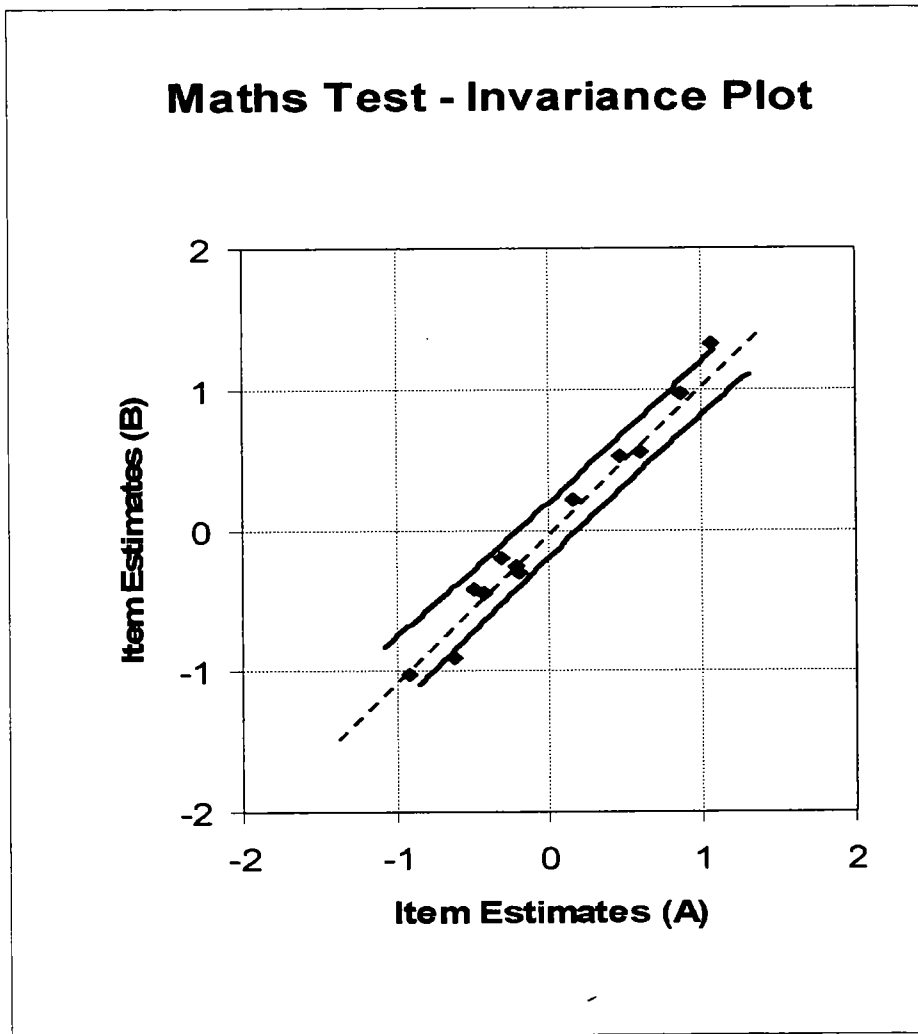
Figure 4.1.5 is the invariance plot as suggested by Wright and Masters (1982) and by Bond and Fox (2001). It is scatter diagram of item measures from subtest B against item measures from subtest A together with the 95% confidence limits based on the errors in the two calibrations.

The dotted line, identity line (Wright and Masters, 1982, p. 115), going through the points represents the exact modelled relation between the two sets of item estimates if they remained completely invariant under perfectly precise (i.e error free) measurement conditions.

The points are closely scattered around the identity line (correlation coefficient between the two sets of measures is 0.989), with only two items lying outside the C.I., and that is a good indication that invariance holds. (See the test after the figure)

The two already mentioned items, 11 and 1, are the ones outside the 95% confidence limits in figure 4.1.5.

Figure 4.1.5 Invariance plot for the maths test



Testing whether 2 items (out of 12) outside the 95% C.I. is unexpected ($p < 0.05$)

In a binomial situation where one has 12 items, each with $P(\text{lying outside the C.I.}) = 0.05$, the expected number of items lying outside the C.I. is 0.6.

Let X = number of items outside the 95% C.I.

$H_0: p = 0.05$ (Under $H_0: X \sim \text{Bin}(12, 0.05)$)

$H_1: p > 0.05$

$P(X \geq 2) = 0.12 \gg 0.05$, therefore we cannot reject H_0 .

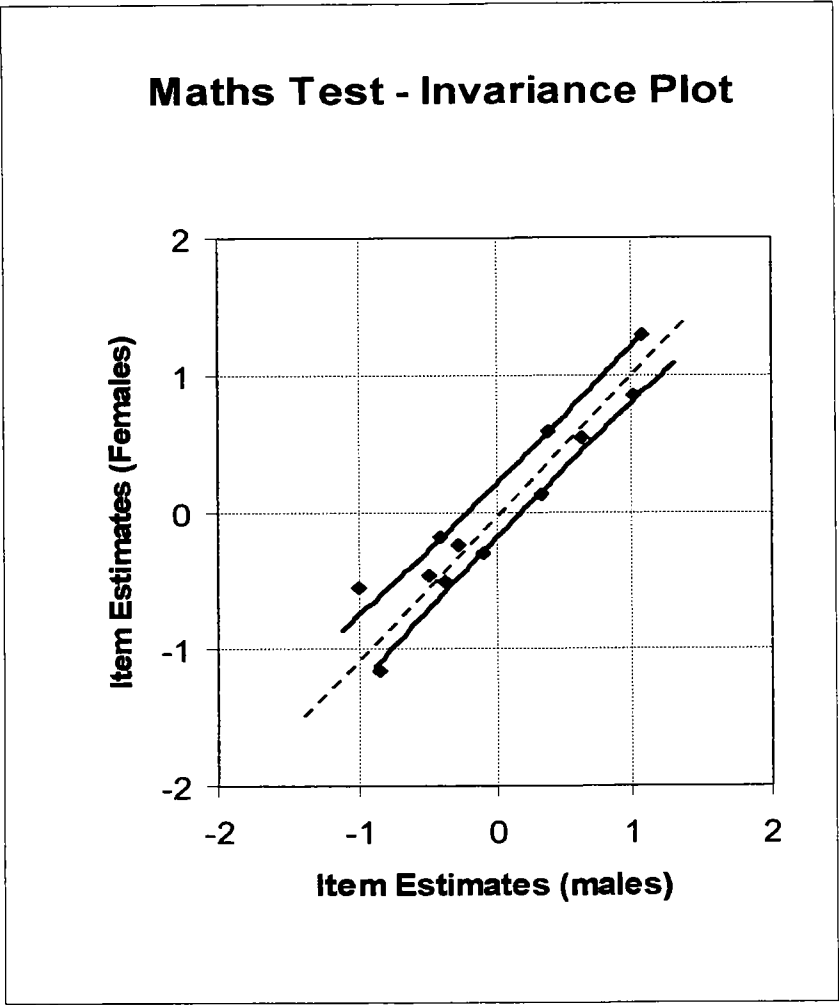
Conclusion: Two points outside the 95% C.I. is not a highly unlikely event if one has 12 items.

(b) Split of the data by gender

In this case the data was split into two groups based on gender. The two groups had sizes 267 (males) and 305 (females).

Figure 4.1.7 below shows the invariance plot for the item estimates from these two subsets.

Figure 4.1.7 Invariance plot for the maths test(by gender)



The points are again closely scattered around the identity line, and again with only 2 out of the 12 items (approximately 16.7% of the items) clearly outside the confidence limits, and that is a good indication that invariance holds. Also, the correlation coefficient is 0.945, also highly significant.

These results support the property of invariance of the Rasch model. That is, when the Rasch model governs measurement one can free item difficulty estimation from the characteristics of persons in the calibration sample. This invariance of item calibrations across groups implies that the construct measured by the instrument has the same meaning to the groups.

All of the above evidence collected in the validation study of this maths test, together with the good fit of the test data to the Rasch model support the hypothesis of a high degree of validity.

4.1.3 Misfitting students

Misfitting students were identified using appropriate cut-off scores for the infit and outfit statistics (1.3 for both). The numbers and proportions of misfitting students are presented, together with comparisons of equivalent proportions from a simulation study.

Following the calibration of the test, misfitting students were identified using cut-off scores for the infit and outfit mean squares of 1.3.

Table 4.1.10 shows the number of students identified as misfitting by the two indices as well as the total number.

Table 4.1.10 Misfit (infit) * Misfit (outfit) Crosstabulation

		Misfit (outfit)		Total
		Fitting	Misfitting	
Misfit (infit)	Fitting	383	50	433
	Misfitting	53	86	139
Total		436	136	572

The number of students identified as misfitting by the outfit statistic was 136 (23.8%) and by the infit statistic was 139 (24.3%), whereas 86 students were identified by both, giving a total of 189 (33%) misfitting students.

A simulation study was carried out. WINSTEPS (Linacre, 2005) provides users the opportunity to use the estimated person, item and structure measures to simulate a Rasch-fitting data set equivalent to the raw data. This can be used to investigate the stability of measures and distribution of fit statistics.

The infit mean square calculated for this Rasch-fitting data set identified 18.2% misfitting students (infit > 1.3) i.e. 104 cases and the outfit mean square 19.4 % (outfit > 1.3) i.e. 111 cases. These two proportions were slightly lower than the proportions found in the empirical data. Simulated data are always expected to fit the Rasch model better and the discrepancy was not great.

The infit and outfit mean square statistics have been shown to follow a Chi-square distribution ($d.f = 1$) with expected value of 1. Even when the data fit the Rasch model perfectly, whatever cut-off scores for identifying misfitting examinees are used (1.2, 1.3, 1.4 or higher) there will be a proportion of examinees with mean square statistics greater than the cut-off score, thus labelled misfitting. The higher the cut-off score the lower the proportion of misfitting students. In other words, whatever cut-off score is used the Rasch model expects a proportion of examinees to have aberrant responses.

The results of the simulation study, show similar proportions with the results from the analyses of the test data, strengthening the belief that the test data collected in phase 1 fit the Rasch model reasonably well.

4.1.4 Test Anxiety Inventory (TAI)

The sample

The TAI was administered to 470 students out of the 572 who took the test (206 males and 264 females). There were two reasons why the sample of students answering the TAI was smaller than the original sample. Those were:

- One teacher who taught two of the classes (the teacher in Limassol 3 school) did not want to administer the TAI to her 44 students and
- 54 students were either absent when the TAI was administered or did not want to complete it.

4.1.5 Validity of the TAI

For the validation study of the test, the following evidence was collected and presented below:

- *Comparisons of TAI results with published analyses in the TAI manual*
- *Factor analysis*
- *Correlations of TAI scores with test scores*

Comparisons with published analyses

The scores of the emotionality and worry factors were calculated using the instructions given in the TAI manual. The scores of the 8 items indicated in the manual as measuring the Worry Factor (items 3, 4, 5, 6, 7, 14, 17 and 20) and the 8 items measuring the Emotionality Factor (items 2, 8, 9, 10, 11, 15, 16, 18) were added to give the score of each factor. Finally, the scores on the 4 remaining items were added to the two factor scores giving the total anxiety scores.

Table 4.1.11 Shows comparisons of the published analyses of the TAI with analyses carried out on the data collected in phase 1.

Table 4.1.11 Published and observed analyses

	High school students (published analyses)		High school students (phase 1 data)		Test for difference between the means (t – values)	
	Male	Female	Male	Female	Male	Female
N	527	591	206	264		
TAI Mean	40.87	45.72	42.36	47.77	0.64	2.15 *
S.D.	12.77	13.63	13.41	12.49		
Alpha	0.92	0.93	0.922	0.923		
Emotionality Mean	16.61	18.91	16.79	20.58	0.38	3.90 **
S.D.	5.47	5.88	5.97	5.74		
Alpha	0.90	0.91	0.880	0.878		
Worry Mean	15.60	17.06	16.66	17.62	2.28 *	1.4
S.D.	5.33	5.76	5.83	5.22		
Alpha	0.86	0.89	0.820	0.816		

* = significant at the 5% level

** = significant at the 1% level

Comparing the results of the published analyses (data collected from high school students in the United States), with the analyses of the data collected in phase 1 (from high school students in Cyprus), it is obvious that these are very similar.

There are significant differences between male high school students only in the mean scores on the Worry subscale ($t = 2.28$, $p = < 0.013$), with the Cypriots scoring significantly higher.

Also, there are significant differences between female high school students on the mean scores on the Emotionality subscale ($t = 3.90$, $p = 0.000$) and the total score on the TAI ($t = 2.15$, $p = 0.016$) with the Cypriots scoring significantly higher.

On the other hand, the variations in the data are almost identical (equal standard deviations) and so are the reliability estimates, the alpha coefficients.

Also, all the alpha coefficients are high indicating that the students' responses are very consistent, and items measure the same or very similar traits.

Factor Analysis

The next three tables, 4.1.12, 4.1.13 and 4.1.14 show the results of factor analysis on the TAI data set. The factors were highly correlated therefore the Principal Axis Factoring method of extraction was used, followed by rotation with the Direct Oblimin method.

Table 4.1.12 Total Variance Explained

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings(a)
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	8.390	41.951	41.951	7.852	39.259	39.259	7.510
2	1.500	7.498	49.449	.910	4.549	43.809	5.263
3	.922	4.610	54.058				
4	.842	4.212	58.270				
5	.782	3.911	62.181				
6	.777	3.884	66.064				
7	.688	3.440	69.504				
8	.677	3.386	72.890				
9	.616	3.078	75.968				
10	.563	2.813	78.780				
11	.522	2.611	81.392				
12	.491	2.457	83.848				
13	.488	2.438	86.286				
14	.466	2.328	88.614				
15	.445	2.225	90.839				
16	.422	2.112	92.951				
17	.420	2.102	95.054				
18	.345	1.727	96.781				
19	.325	1.625	98.405				
20	.319	1.595	100.000				

Extraction Method: Principal Axis Factoring.
a When factors are correlated, sums of squared loadings cannot be added to obtain a total variance.

Table 4.1.13 The items and their correlations with the two factors extracted

Statements	Factors	
	Emot.	Worry
1. I feel confident and relaxed while taking exams	,604	,470
<u>2. While taking examinations I have an uneasy, upset feeling</u>	<u>,697</u>	,410
3. Thinking about my grade in a course interferes with my work in tests	,458	,562
4. I freeze up on important exams	,649	,486
5. During exams I find myself thinking whether I will ever get through school	,308	,500
6. The harder I work at taking a test, the more confused I get	,308	,572
7. Thoughts of doing poorly interfere with my concentration on tests.	,572	,624
<u>8. I feel very jittery when taking an important test.</u>	<u>,728</u>	,446
<u>9. Even when I'm well prepared for a test, I feel very nervous about it</u>	<u>,648</u>	,412
<u>10. I start feeling very uneasy just before getting a test paper back</u>	<u>,606</u>	,463
<u>11. During tests I feel very tense.</u>	<u>,783</u>	,523
12. I wish examinations did not bother me so much	,569	,474
13. During important tests I am so tense that my stomach gets upset.	,689	,423
14. I seem to defeat myself while working on important tests	,425	,651
<u>15. I feel very panicky when I take an important test</u>	<u>,781</u>	,521
<u>16. I worry a great deal before taking an important examination</u>	<u>,744</u>	,394
17. During tests I find myself thinking about the consequences of failing	,556	,668
<u>18. I feel my heart beating very fast during important tests</u>	<u>,691</u>	,400
19. After an exam is over I try to stop worrying about it but I can't	,558	,494
20. During examinations I get so nervous that I forget facts I really know	,585	,628

Items in bold are the items on the worry subscale (according to the TAI manual) and items in bold and underlined are the ones on the emotionality subscale. The 4 items that

are neither bold nor underlined are the ones whose scores, combined with the scores of the other 16 items make the total anxiety score.

Table 4.1.14 Factor Correlation Matrix

Factor	1	2
1	1,000	,636
2	,636	1,000

All items load significantly on both factors (r well above 0.3) as expected because of the high correlation (0.636) between the two factors.

Items 2, 8, 9, 10, 11, 15, 16, 18 have much higher loadings on factor 1 therefore we can conclude that factor 1 in this dataset is the emotionality factor.

With the exception of item 4, the other worry items have much higher loadings on factor 2 therefore we can conclude that factor 2 in this dataset is the worry factor. Item 4 may have a higher loading on factor 1 (0.659), however it still loads significantly (0.495) with factor 2.

Furthermore, the correlation coefficient between the scores on the two factors as they appear in the data set (0.676) is very similar to the correlation coefficient between the two factors extracted by factor analysis (0.636) strengthening the belief that the two factors extracted are indeed the factors of emotionality and worry.

Correlations of TAI scores with test scores

The correlations of the test scores with the emotionality score, the worry score and the total anxiety score were also calculated (table 4.1.15)

All three have negative, statistically significant correlations with the test score. The strongest correlation is between the test score and the worry factor (-0.38).

These results are similar to the published analyses for the validity of the test where the TAI scores of high school students were compared with an IQ test and with their grade point average (GPA). In all the analyses the correlations were negative, some significant and some not, but in all cases the Worry factor had the strongest correlation.

Table4.1.15 Correlations of TAI scores with the test scores

	Anxiety score	Emotionality score	Worry score
Test score	- 0.263 **	- 0.119 **	- 0.381 **

** Correlation is significant at the 0.01 level (2-tailed).

4.1.6 TAI calibrations

The Rasch model was used for the calibrations. The first calibration on the full dataset revealed two slightly misfitting items and six badly misfitting students (outfit > 3.0).

The six students were removed and a second calibration was performed, improving the outfit and infit values of the two misfitting items. Those items were retained in the dataset (the reasons for not removing the items are explained).

The item statistics from the second calibration were then used for the final calibration, to get the students statistics.

Item-person maps are presented to show how well the items are targeted for the population of students and finally the students are divided into groups according to their anxiety estimates for investigating later on whether anxiety is associated with misfit.

First calibration

The first calibration revealed two slightly misfitting items, item 5 (During exams I find myself thinking whether I will ever get through school. Outfit = 2.00, infit = 1.84) and item 6 (The harder I work at taking a test, the more confused I get. Outfit = 1.64, infit = 1.50). The cut-off score used for both the statistics is 1.5 as suggested by Wright et al. (1994).

Table 4.1.16 shows the item statistics in misfit order.

Table 4.1.16 ITEM STATISTICS: MISFIT ORDER

ENTRY	RAW					INFIT	OUTFIT	PTMEA	
NUMBER	SCORE	COUNT	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	ZSTD	CORR.
5	784	466	1.11	.07	1.84	9.6	2.00	8.9	A .43
6	874	466	.70	.06	1.50	6.7	1.64	7.0	B .45
12	1160	466	-.37	.06	1.15	2.4	1.25	3.6	C .59
14	1097	466	-.15	.06	1.18	2.9	1.19	2.8	D .55
3	1122	465	-.24	.06	1.15	2.5	1.18	2.7	E .56
13	873	466	.70	.07	1.05	.8	.96	-.6	F .64
9	1037	466	.07	.06	1.01	.2	1.05	.7	G .62
4	1040	466	.06	.06	1.02	.4	.98	-.2	H .65
10	1275	466	-.78	.06	.99	-.2	1.01	.2	I .62
19	886	466	.65	.06	1.01	.1	1.01	.1	J .59
1	1281	466	-.80	.06	.83	-3.0	.98	-.2	j .63
18	1009	466	.17	.06	.98	-.2	.98	-.3	i .64
17	1034	466	.08	.06	.97	-.6	.97	-.4	h .63
20	1042	466	.05	.06	.95	-.8	.90	-1.6	g .65
8	993	466	.23	.06	.90	-1.6	.87	-1.9	f .67
2	1275	466	-.78	.06	.86	-2.5	.89	-1.6	e .66
7	1101	465	-.17	.06	.87	-2.3	.86	-2.3	d .65
16	1217	465	-.58	.06	.78	-4.1	.75	-4.1	c .68
15	1017	466	.14	.06	.70	-5.4	.68	-5.4	b .72
11	1080	466	-.09	.06	.58	-8.2	.62	-6.6	a .73
MEAN	1060.	466.	.00	.06	1.02	-.2	1.04	.0	
S.D.	136.	0.	.51	.00	.27	3.8	.31	3.6	

Table 4.1.17 shows the top part of the person statistics in misfit order (From the most misfitting students to ones with infit and/or outfit = 2.0). Six students were identified as badly misfitting (outfit > 3.0). Furthermore table 4.1.18 shows the response strings of these six students and 4 of those had unexpected responses to item 5 and one to item 6. Students in this table are in misfit order (most to least misfitting) and items in difficulty order (Easier to most difficult).

Table 4.1.17 Student statistics in misfit order

ENTRY	RAW					INFIT	OUTFIT	PTMEA	
NUMBER	SCORE	COUNT	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	ZSTD	CORR.
262	27	20	-2.13	.41	3.21	3.7	6.60	6.2	A-.50
325	31	20	-1.57	.34	2.68	3.5	3.81	4.8	B-.39
311	77	20	3.01	.59	1.76	1.2	3.54	2.5	C-.31
466	75	20	2.47	.47	2.71	2.6	3.47	3.1	D-.08
339	21	20	-4.24	1.01	1.06	.4	3.32	1.7	E-.50
460	36	20	-1.05	.30	2.60	3.8	3.24	4.7	F-.17
313	36	20	-1.05	.30	2.51	3.6	2.73	3.9	G-.05
310	58	20	.59	.27	2.65	4.4	2.67	4.3	H-.18
307	32	20	-1.45	.33	2.17	2.8	2.57	3.3	I-.02
230	39	20	-.79	.29	2.39	3.6	2.51	3.7	J .14
113	47	20	-.19	.27	2.34	3.7	2.37	3.8	K-.22
245	41	20	-.63	.28	2.23	3.3	2.30	3.4	L-.09
232	76	20	2.70	.52	2.28	2.0	1.48	.9	M .31
356	35	20	-1.15	.31	1.36	1.1	2.10	2.7	N .21
305	61	20	.82	.28	2.09	3.1	2.03	2.9	O .06
515	78	20	3.43	.72	1.99	1.3	2.06	1.3	P .03
216	56	20	.45	.27	2.05	3.1	2.00	3.0	Q .32
239	35	19	-.95	.30	1.81	2.2	2.04	2.6	R .19
223	40	18	-.35	.29	2.01	2.8	2.01	2.8	S .40

Table 4.1.18 Most misfitting response strings

student	OUTMNSQ		item	
			1 11	112 111 11
			10262374104975889365	
high-----				
262 2404	6.60	A		44
325 3109	3.81	B	34.....	4
311 2616	3.54	C	2.....3.....	
466 4404	3.47	D	1.....2....	
339 3123	3.32	E		2
460 4120	3.24	F	4.....4...4	

Second calibration

These 6 students (1.3% of the students) were removed from the data set and a second calibration was run, improving the infit and outfit statistics for items 5 and 6. Although the item fit statistics are still over the cut-off score of 1.5, now they are only just above and since the internal consistency of the test is very high, and this test is simply a questionnaire where very accurate estimates of trait measure are not really necessary, the two items were retained in the instrument.

Table 4.1.19 item statistics: misfit order

ENTRY	RAW	COUNT	MEASURE	ERROR	INFIT		OUTFIT		PTMEA	items
NUMBER	SCORE				MNSQ	ZSTD	MNSQ	ZSTD	CORR.	
5	762	460	1.17	.07	1.78	9.0	1.63	6.0	A .45	item 5
6	858	460	.72	.07	1.49	6.5	1.57	6.4	B .45	item 6
3	1109	459	-.25	.06	1.17	2.7	1.21	3.0	C .55	item 3
14	1082	460	-.14	.06	1.18	2.9	1.19	2.8	D .55	item 14
12	1151	460	-.39	.06	1.13	2.1	1.15	2.3	E .60	item 12
13	860	460	.71	.07	1.07	1.0	.97	-.4	F .63	item 13
9	1024	460	.07	.06	1.02	.3	1.05	.8	G .62	item 9
4	1026	460	.06	.06	1.04	.6	1.00	.0	H .64	item 4
10	1263	460	-.80	.06	1.00	.0	1.03	.4	I .61	item 10
19	874	460	.65	.06	1.02	.3	1.02	.4	J .57	item 19
18	997	460	.17	.06	1.00	.0	.99	-.1	j .63	item 18
17	1022	460	.07	.06	.98	-.4	.99	-.1	i .62	item 17
20	1030	460	.04	.06	.97	-.6	.91	-1.4	h .64	item 20
2	1262	460	-.79	.06	.87	-2.2	.91	-1.4	g .65	item 2
8	980	460	.23	.06	.88	-1.9	.84	-2.5	f .68	item 8
1	1269	460	-.82	.06	.83	-3.0	.87	-2.1	e .63	item 1
7	1084	459	-.16	.06	.86	-2.5	.85	-2.4	d .64	item 7
16	1205	459	-.59	.06	.78	-3.9	.76	-4.0	c .67	item 16
15	1005	460	.14	.06	.71	-5.2	.69	-5.2	b .72	item 15
11	1067	460	-.09	.06	.59	-8.0	.63	-6.5	a .73	item 11
MEAN	1046.	460.	.00	.06	1.02	-.1	1.01	-.2		
S.D.	137.	0.	.52	.00	.26	3.7	.24	3.2		

Third and final calibration

Finally, the estimates for the 20 items were used (items were anchored) for the final calibration which included all the 474 students.

A summary of the results of the Rasch analysis from the final calibration is given in table 4.1.20

Table 4.1.20 Summary of the results of the Rasch analysis for the TAI

	N	Estimate of mean (SD)	Range	Reliab.	Separ. Index	Infit msq mean (SD)	Outfit msq mean (SD)
Examinees	470	-0.38 (1.12)	-4.26 to 3.45	0.90	3.03	1.02 (0.45)	1.05(0.57)
Items	20	0.0 (0.52)	-0.82 to 1.17	0.98	7.98	1.03 (0.28)	1.05(0.32)

The range of student measures was from -4.26 to 3.45 (excluding the maximum and minimum scores), with a mean of -0.38 (SD = 1.12). The reliability of student estimates was 0.90 and the separation index was 3.03. This separation index indicates that the instrument identifies 4 statistically distinct strata of student anxiety levels.

The item estimates ranged from -0.82 to 1.17 and the reliability index was 0.98 (separation index = 7.98).

Figure 4.1.8 shows the item-student map. One can see that the test items are well targeted for students with anxiety measure from half a standard deviation below the mean to one and a half standard deviations above the mean measure. That is, the test items are well targeted for about 63% of the distribution of students' measures.

Figure 4.1.8 STUDENTS MAP OF items

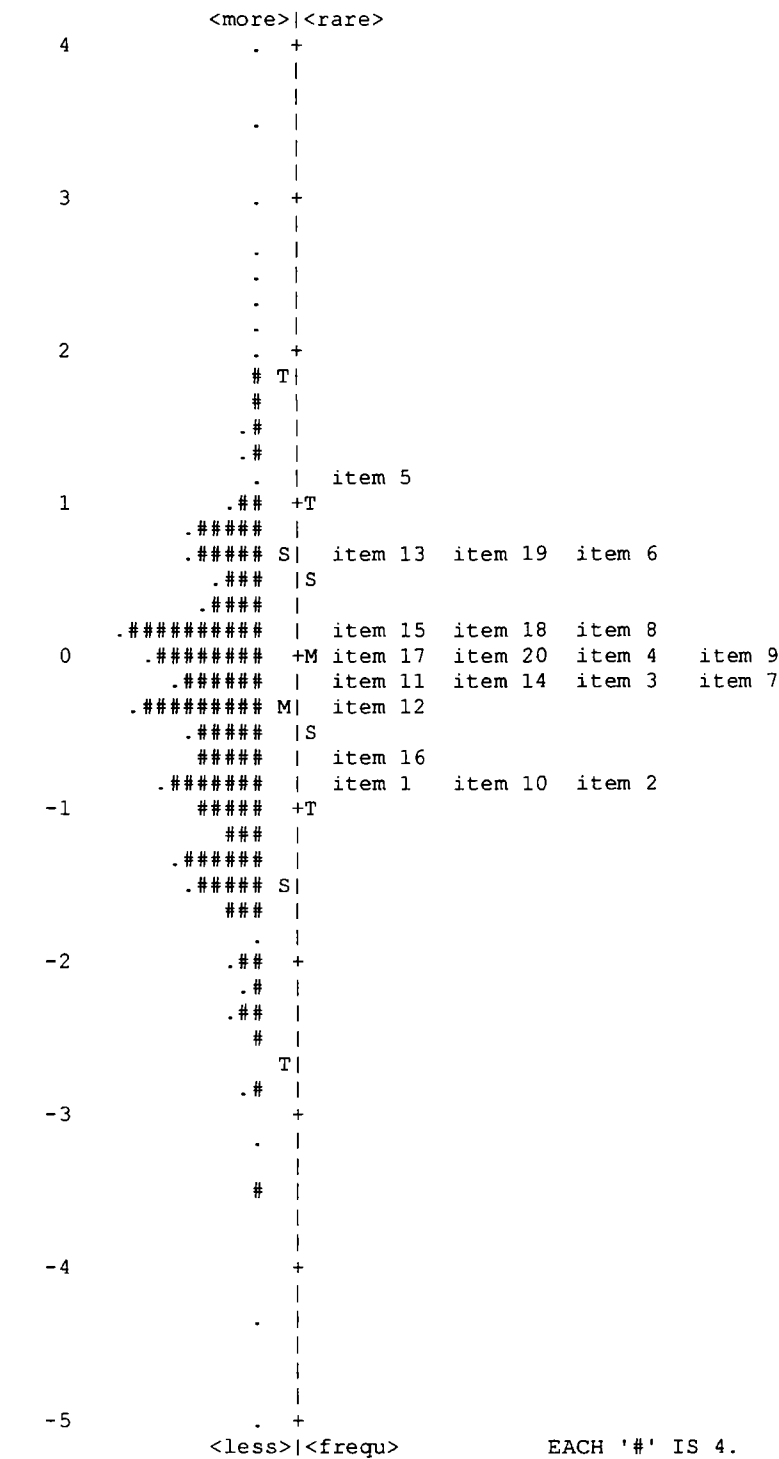


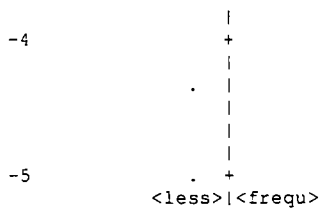
Figure 4.1.9 is another item – student map, this time with all the thresholds for the possible scores (1 to 4) for each item. It is obvious that the various steps of the questions are well targeted for a wider range of abilities, from 2 standard deviations below the overall mean measure to the top of the measures' distribution.

Figure 4.1.9 students MAP OF items

```

4      <more>|
        . +
          |
          |
        . |
          |
          |
3      . +
          | item 5 .4
        . |
        . |
        . | item 13.4
          | item 6
          | item 19.4
2      . |
        . +
        # T| item 18.4
          | item 8
        # | item 15.4
          | item 17
          | item 20
          | item 4
          | item 9
        .# | item 11.4
          | item 14
          | item 7
        .# | item 3 .4
        . | item 5 .3
        .## +T item 12.4
        .##### | item 16.4
          | item 1 .4
          | item 10
          | item 2
        .##### S| item 13.3
          | item 19
        .### |S
        .##### | item 8 .3
        .##### | item 15.3
        .##### | item 17
          | item 18
          | item 4
          | item 9
0      .##### +M item 11.3
          | item 20
        .##### | item 14.3
          | item 3
          | item 7
        .##### M| item 12.3
        .##### |S item 5 .2
        ##### | item 16.3
        ##### |
        .##### | item 13.2
          | item 6
          | item 1 .3
          | item 10
          | item 2
-1     ##### +T item 19.2
        ### |
        .##### | item 8 .2
        .##### S| item 15.2
          | item 17
          | item 18
          | item 20
          | item 4
          | item 9
        ### | item 11.2
        . | item 14.2
          | item 3
          | item 7
-2     .## + item 12.2
        .# | item 16.2
        .## | item 2 .2
        # | item 1 .2
          | item 10
          T|
        .# |
-3     . +
          |
          |
        # |
          |

```



Different anxiety groups

The range of abilities (anxiety measures) was divided into three different groups, the low, medium and top anxiety groups using three different cut-off scores.

A similar procedure was used for categorising the students as was used for the maths test. First, the range of ‘abilities’ was divided into 3 groups using the 30th (measure of -0.885) and 70th (measure of 0.165) percentiles. The lowest 30% of the distribution was labelled ‘Low 30% Anxiety’ group, the middle 40% ‘Medium Anxiety’ group and the top 30% ‘Top 30% Anxiety’ group.

Second, the range of ‘abilities’ was divided into 3 groups using the 20th (measure of -1.356) and 80th (measure of 0.525) percentiles. The lowest 20% of the distribution was labelled ‘Low 20% Anxiety’ group, the middle 60% ‘Medium Anxiety’ group and the top 20% ‘Top 20% Anxiety’ group.

Third, the range of ‘abilities’ was divided into 3 groups using the 10th (measure of -1.823) and 90th (measure of 0.919) percentiles. The lowest 10% of the distribution was labelled ‘Low 10% Anxiety’ group, the middle 80% ‘Medium Anxiety’ group and the top 10% ‘Top 10% Anxiety’ group.

Given the high negative correlation between the scores on the worry factor and the test scores, reported both in the TAI manual and in the data in this study, the range of ‘worry scores’ was again divided into 3 groups using the same three cut-off percentiles as in the anxiety measures. However instead of the measures, the worry raw scores were used instead.

Therefore, the data was divided into ‘Low 30% Worry’, ‘Medium Worry’ and ‘Top 30% Worry’ group and similarly for the 20 and 80% and 10 and 90%.

4.1.7 Misfitting students

Misfitting students were identified using appropriate cut-off scores for the infit and outfit statistics. The numbers and proportions of misfitting students are presented. Finally, a chi-square (contingency tables) test is performed to investigate possible association between misfit in the maths test and misfit in the TAI.

Following the calibration of the TAI, misfitting students were identified using cut-off scores for the infit and outfit mean squares of 1.5. Table 4.1.21 shows the number of students identified as misfitting by the two indices as well as the total number.

Table 4.1.21 Misfit (infit) * Misfit (outfit) Crosstabulation for the TAI

		Misfit (outfit)		Total
		Fitting	Misfitting	
Misfit (infit)	Fitting	397	12	409
	Misfitting	7	54	61
Total		404	66	470

The number of students identified as misfitting by the outfit statistic was 66 (14.0%) and by the infit statistic was 61 (13.0%), whereas 54 students were identified by both giving a total of 73 (15.5%) misfitting students.

Fifty four out of the 73 (74%) misfitting students were identified by both the person fit statistics (as opposed to 45% in the test). That means that these students had unexpected responses on both on-target and extreme items, based on the distance of the items from their ability on the item-student map. This probably shows that they answered the questionnaire without too much concentration or care, especially since they knew that the results of the questionnaire would only be used for the purposes of this study and had no effect on their school grades or overall assessment.

4.1.8 Assessment of Attention Deficit Hyperactivity Disorder (ADHD) characteristics

Towards the end of the academic year the mathematics teachers were asked to rate the severity of ADHD symptoms of all their students, as observed in the classroom setting, using an 18-item rating scale that was based on the diagnostic criteria of ADHD (American Psychiatric Association, 1994) contained in the Diagnostic and Statistical Manual of Mental Disorders Version 4 (DSM IV).

This instrument was based on dichotomous items on which teachers were asked to consider a criterion met if the behaviour had persisted for at least six months and it was considerably more frequent than that of most other students of the same developmental level.

Of the 12 teachers involved, 2 did not want to complete the scale (each teaching two classes) and one completed it for only one of the two classes he was teaching, leaving 20 classes and a total of 441 (out of the original 572) students composing this sample.

Reliability and Validity of the instrument

For the study of the reliability of the scale Cronbach's alpha was used. Furthermore, the item-total correlations were calculated and used as another indication of a high degree of internal consistency of the test.

For the validation study of the test, the following evidence was collected and presented below:

- *Factor analysis*
- *Correlations of assessments of different teachers in 4 classes.*

Table 4.1.22 shows the alpha coefficient which is a measure of the internal consistency of the test. It is very high (0.953) because all the questionnaires were completed by 10 teachers, who completed them in a very consistent way.

Table 4.1.22 Reliability Statistics

Cronbach's Alpha	N of Items
,953	18

The item-total correlations (measures of the discriminating power of each item) varied from 0.588 to 0.818 and were all highly significant ($p < 0.01$).

Factor analysis performed on the data (using the Principal components analysis method) extracted two factors, as described by the scale in DSM IV, shown in table 4.1.23

Table 4,1,23 Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Var	Cum.%	Total	% of Var	Cum. %	Total	% of Var.	Cum. %
1	10,255	56,974	56,974	10,255	56,974	56,974	6,715	37,305	37,305
2	3,082	17,123	74,098	3,082	17,123	74,098	6,623	36,793	74,098
3	,846	4,699	78,796						
4	,635	3,530	82,326						
5	,469	2,605	84,931						
6	,415	2,308	87,239						
7	,321	1,786	89,025						
8	,291	1,617	90,642						
9	,272	1,513	92,155						
10	,252	1,399	93,554						
11	,196	1,087	94,641						
12	,189	1,052	95,693						
13	,172	,957	96,650						
14	,148	,823	97,473						
15	,134	,747	98,220						
16	,128	,712	98,932						
17	,103	,570	99,502						
18	,090	,498	100,00						

Table 4.1.24 shows the factor loading of the items on the factors (rotation, with varimax)

Table 4.1.24 Rotated Component Matrix

	Component	
	1	2
a1	,855	,111
a2	,891	,212
a3	,690	,400
a4	,908	,218
a5	,907	,124
a6	,858	,158
a7	,536	,478
a8	,694	,413
a9	,842	,249
a10	,369	,774
a11	,299	,759
a12	,409	,782
a13	,551	,653
a14	,207	,751
a15	,298	,830
a16	,051	,888
a17	,109	,896
a18	,100	,894

It is evident in table 4.1.24 that the first 9 items (described in DSM IV as the ones measuring inattention) load significantly on factor 1 and the last 9 items (described in DSM IV as the ones measuring hyperactivity/impulsivity) load significantly on factor 2.

Therefore the results of the factor analysis support the validity of the instrument in that they identify exactly the two factors described in the manual.

Furthermore, in 4 classes, 3 from Limassol 1 and 1 from Limassol 2 (a total of 90 students) the ADHD scale was given also to the Language teachers to assess the behaviour of their students. The number of criteria met by students, as assessed by the language teachers, were compared with the ones from the mathematics teachers' assessments, and were found to be highly correlated. The correlation coefficients were:

$r = 0.73^{**}$ for the Inattention subscale.

$r = 0.63^{**}$ for the Hyperactivity/Impulsivity subscale.

$r = 0.78^{**}$ for the Combined scale.

The assessments of the language and mathematics teachers as to whether students could be considered as displaying ADHD symptoms of anyone of the ADHD subtypes agreed on 75 of the 90 students.

Table 4.1.25 shows the n numbers of students categorized as displaying ADHD symptoms based on the Maths and Language teachers’ assessments.

Table 4.1.25 Number of students displaying ADHD symptoms based on teachers assessments

		Language Teachers		Totals
		ADHD	No ADHD	
Maths Teachers	ADHD	21	9	30
	No ADHD	6	54	60
Totals		27	63	90

Most of the disagreements were because of a small difference in the number of criteria met by the students.

Results of the teachers’ ratings

The proportion of pupils observed by their teachers to display ADHD symptoms in the classroom setting was 30.4% (i.e. 30.4% of the students, based on their teachers’ ratings, were found to meet at least 6 out of the 9 criteria in one, or both, of the subscales in the ADHD scale). The proportions of the three subtypes of ADHD, Predominantly Inattentive, Predominantly Hyperactive/Impulsive and Combined, according to the teachers’ ratings were 21.5% (95 students), 2.0% (9 students) and 6.8% (30 students) respectively.

Table 4.1.26 shows the number of boys (85, 40.7% of the total number of boys) and girls (49, 21.1% of the total number of girls) observed by their teachers to display ADHD symptoms.

There is a highly significant difference ($p = 0.000$) between the proportions of boys observed to display ADHD symptoms and the proportion of girls, with boys having almost double the proportion of girls.

Table 4.1.26 Gender * ADHD as observed by maths teachers Crosstabulation

		ADHD as observed by the maths teachers		
		No ADHD symptoms	ADHD symptoms	Total
Gender	Male	124	85	209
	Female	183	49	232
	Total	307	134	441

Chi-square = 18.921, d.f. = 1, p = 0.000

The ratio of boys to girls observed to display ADHD symptoms (1.93:1) is almost identical to the ratio of 2:1 reported by Barkley and Murphy (1998, pp. 6-7) for adults. However, the proportion of students observed to display ADHD symptoms (30.4%) is much higher than the 3 – 7 % of the childhood population, or the 2 – 5 % of the adult population reported by Barkley and Murphy (1998, pp. 6-7). It is even much higher than the estimated proportion of 8.1% to 17% of primary school children observed by their teachers to display severe ADHD symptoms (Merrell and Tymms, 2001).

Possible reasons for this high proportion could include:

- Teachers in Cyprus have not been familiar with using the DSM IV scales, or with other similar scales, therefore this was a new experience to them.
- These teachers are lyceum teachers whose students are adolescents and the assessment of the behaviour of their students was context specific. In other words, the assessment was with respect to the students’ behaviour only in the maths class.
- According to Rice (1999) adolescence is a human developmental stage where the important goal is independence, and the route to that goal is not an easy one; it involves physical, emotional, social, intellectual and spiritual development. Adolescence has traditionally been viewed as a period of “storm and stress”, (Rice, 1999, p. 1) and teenagers’ behaviour can easily be mistaken, especially by adults as ADHD behaviour.
- Finally, first form students in the lyceums in Cyprus have no options in subject selections. All of them have to take core mathematics for 4 periods a week. Therefore, given the well known weakness of a large proportion of students in mathematics, one can easily mistake these weaknesses and consequent indifference as ADHD symptoms.

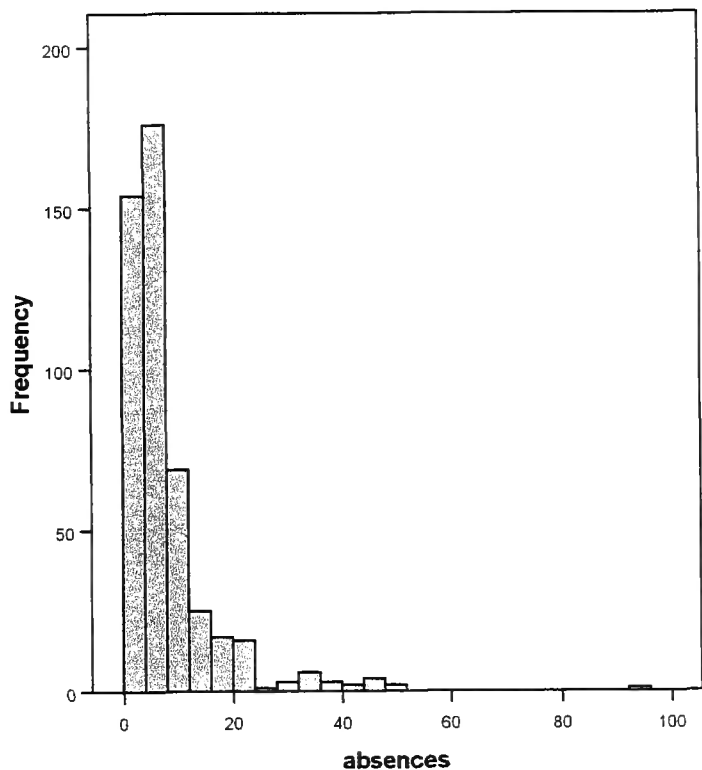
4.1.9 Other factors considered

Factors other than ability, anxiety and ADHD behaviour are also included in the investigation for possible associations with misfit. These factors are explained as to how they were obtained and how most of them were converted into categorical variables. Analyses start with atypical schooling and language competency, where the data was continuous.

Atypical schooling - descriptives

The number of unauthorized absences during the first term of the academic year was used as an indication of atypical schooling. The distribution of the numbers of absences is shown in figure 4.1.10. It is positively skewed with a mean of 7.57 and standard deviation of 9.301.

Figure 4.1.10. Histogram: Distribution of students' absences.



Of those 479 students who answered the specific question, 81% (388 students) had up to and including 10 absences, whereas 6.1% and 1.9% had more than 20 or 40 absences respectively.

Atypical schooling – fit analysis

The next two figures 4.1.11 and 4.1.12 show the scatter diagrams of infit and outfit statistics against the number of absences.

Figure 4.1.11. Scatter diagram of infit vs absences

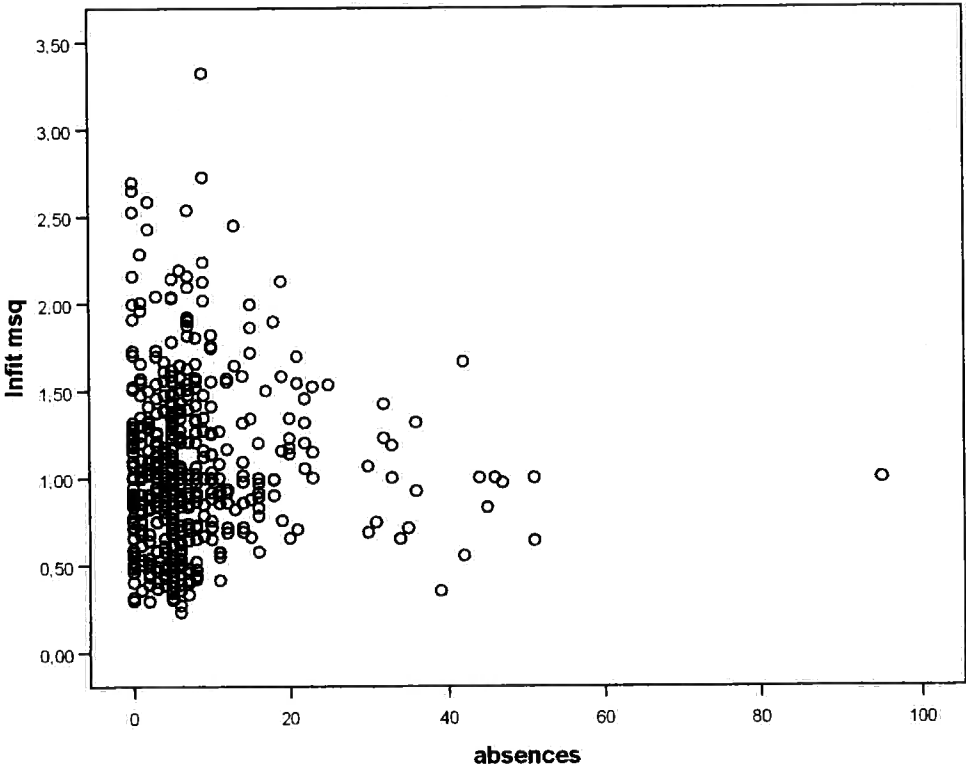
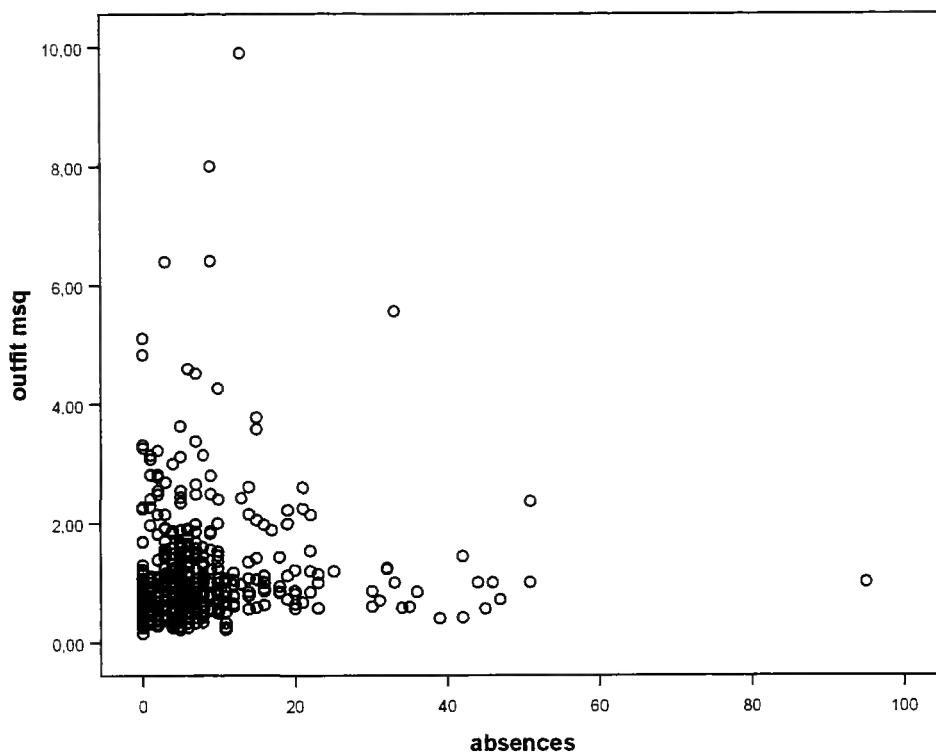


Figure 4.1.12. Scatter diagram of outfit vs absences



The two diagrams indicate that there is no relationship between the infit or outfit and the number of students' unauthorized absences. This finding is strengthened by the correlation coefficients which are: $r = 0.028$ for infit vs absences and $r = 0.041$ for outfit vs absences

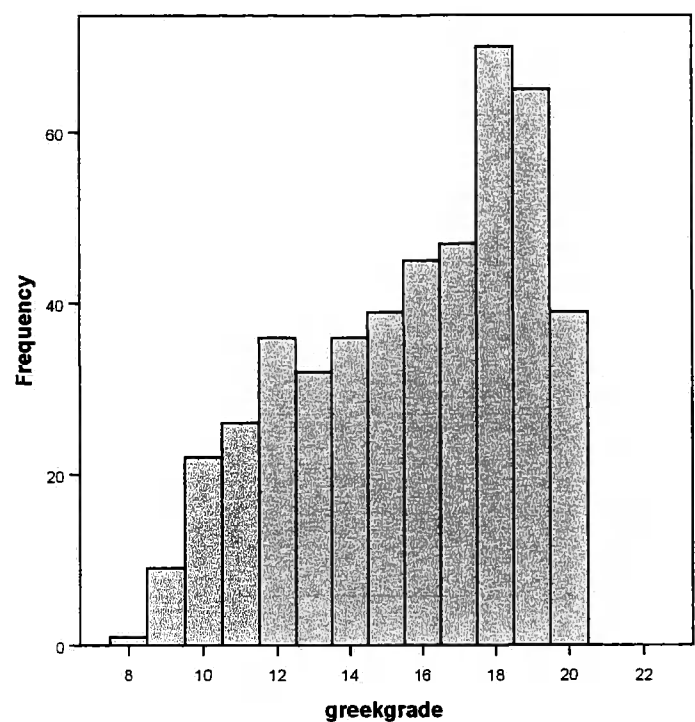
Therefore one can safely conclude that atypical schooling, measured by the number of unauthorized absences, is not a factor affecting misfit. (i.e. there is no indication that students with more absences will have higher infit or outfit values.

Language competency - descriptives

The first term grade in Greek language of each student was used as a measure of language competency. The grades of students in public schools vary from 1 to 20. However it is common practice to use 8 as the minimum grade. The language grade of 105 students could not be obtained, leaving only 467 of the 572 students.

The histogram in figure 4.1.13 below shows the distribution of the language grades. The distribution is negatively skewed, with a mean of 15.7 and standard deviation of 3.1.

Figure 4.1.13. Histogram: Distribution of language grades



Language competency – fit analysis

The next two figures 4.1.14 and 4.1.15 show the scatter diagrams of infit and outfit statistics against the language grades.

Figure 4.1.14 Scatter diagram of infit vs language grades

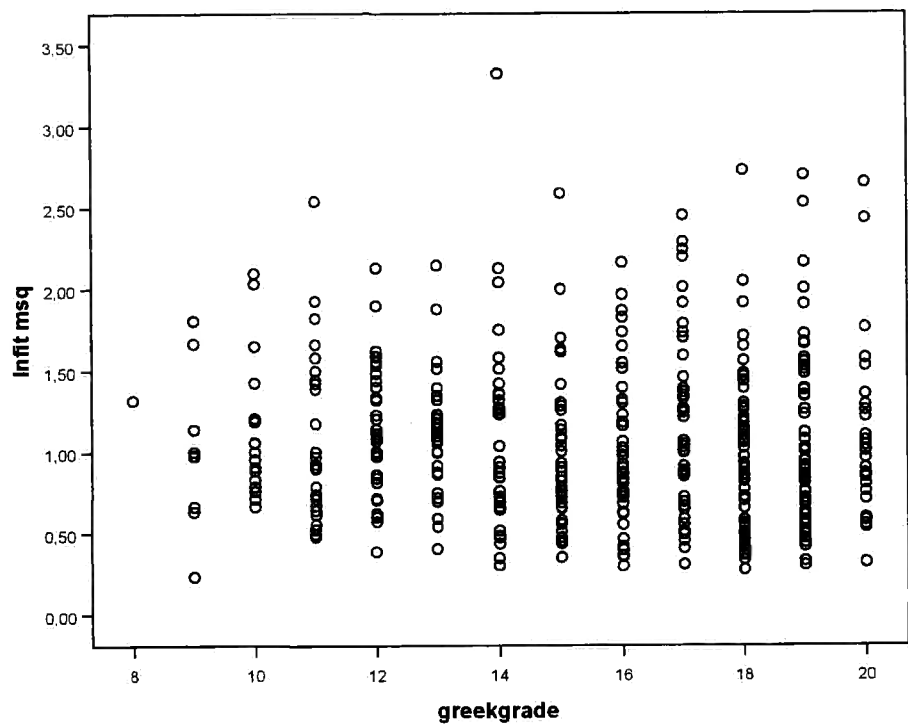
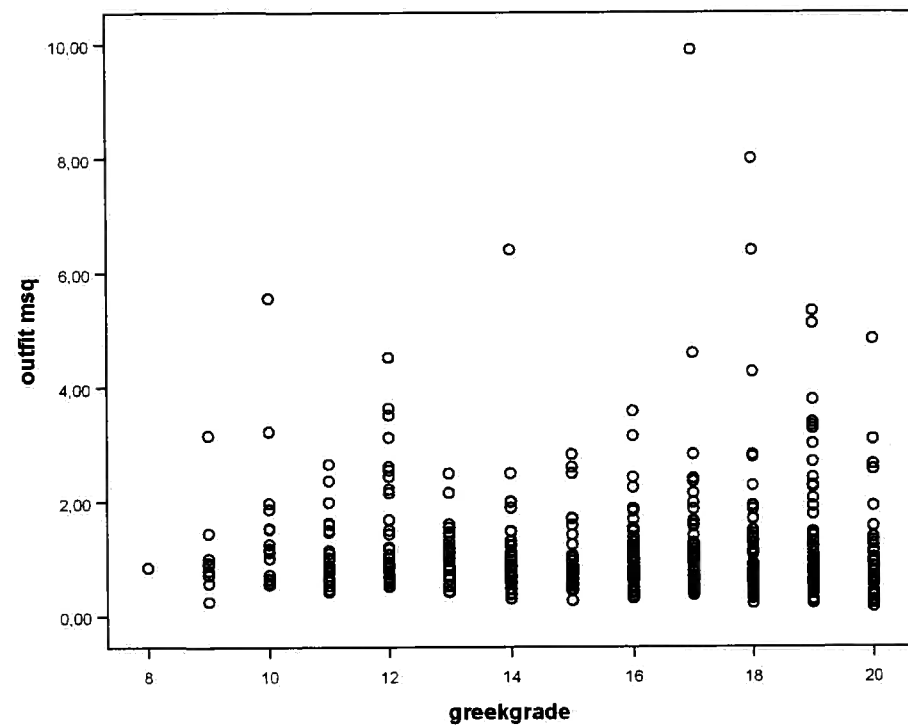


Figure 4.1.15 Scatter diagram of outfit vs language grades



The two diagrams indicate that there is no real relationship between the infit or outfit and the language grade.

This finding is strengthened by the correlation coefficients between these variables which in this case are negative, but not significant. They are:

$r = - 0.048$ for infit vs grade and $r = - 0,010$ for outfit vs grade

Therefore one can safely conclude that language competency, measured by the first term language grade, is not a factor affecting misfit. (i.e. there is no indication that students who are less competent in language will have higher infit or outfit values).

Categorical variables

All the categorical variables are presented below with explanations as to how they were categorised and the number of students in each category. Fit analysis follows in the next section.

Item order

Although all the tests had the same items, those were given in two different orders, A and B. The 12 items of the test were laid out in 4 pages. In order B the items in each of the 4 pages were exactly the same as the items in order A but in reverse order. Out of the 572 students 290 (50.7%) answered item order A and 282 (49.3%) item order B.

Interest in maths

The maths teachers were asked to assess the interest their students show in the subject, using a 3-point Likert scale where 1 = none, 2 = sometimes interested and 3 = always interested. One of the values was missing, leaving a total of 571 students.

Table 4.1.27 shows the frequencies and percentages of students in each group.

Table 4.1.27 Interest in maths

	Frequency
None	150
Sometimes interested	223
Always interested	198
Total	571

Based on the assessment of the maths teachers 26.3% of the students showed no interest in the subject, 39.1% were sometimes interested and 34.7% were always interested.

Private tuition in mathematics

A very large proportion of students (58.2%) take private tuition in mathematics. Students had to complete a very short questionnaire attached to the TAI asking them, among other things, whether they were taking private tuition in mathematics. Out of the 572 students 469 answered that question stating yes or no, leaving 103 missing values in the data.

Table 4.1.28 shows the proportions of boys (52.9%) and girls (62.4%) taking private tuition in mathematics. There are significant differences between students in the two genders with more girls than boys taking the private lessons.

*Table 4.1.28 Private tuition * Gender Crosstabulation*

		Gender		Total
		Male	Female	
Do you take private tuition in maths?	Yes	109 52.9%	164 62.4%	273 58.2%
	No	97 47.1%	99 37.6%	196 41.8%
Total		206	263	469

Chi-square Likelihood ratio = 4.233 (p = 0.04)

Study time

Students were asked to state in the short questionnaire attached to the TAI, how much time on average they usually spend studying mathematics every day. Students spending more than 30 minutes (66 students, 14.1%) were labeled ‘hard workers’ and those spending 10 minutes or less (71 students, 15.2%) ‘lax workers’. The remaining 330 students, spending between 10 and 30 minutes were labeled ‘regular workers’. 105 students did not answer that question. These cut-off times were decided after discussion with the maths teachers in Limassol 1.

Table 4.1.29 shows the number and percentage of students in each of the groups.

Table 4.1.29 Study time groups

	Frequency
Inconsistent workers	71
Regular workers	330
Hard workers	66
Total	467

Preference for mathematics

Students were also asked to state, in the short questionnaire attached to the TAI, whether mathematics was one of their favorite subjects. 103 students did not answer this question. Out of the remaining students 197 (42%) answered ‘Yes’, 272 (58%) ‘No’. There were no gender differences in the preference for maths, with 40.7% for the females and 43.7% for the males answering this question with a ‘Yes’.

Teaching periods spent on revision before the test

Teachers were asked to complete a short questionnaire analyzing the periods spent on revision before the test. Two teachers did not complete the questionnaire leaving 88 missing values.

The total number of periods varied from 1 to 3. Out of the 484 students (572 – 88 missing values) 189 (39%) had 1 period, 251 (51.9%) had 2 periods and 44 (9.1%) had 3 periods for revision.

Log-linear analysis

A brief description of the method of log-linear analysis is given below. A more detailed discussion of log-linear analysis is given in the appendices.

Log-linear analysis is a multivariate extension of the chi-square contingency tables. It is a goodness-of-fit test that can be used for contingency tables with two or more categorical variables. It allows one to test all the effects (main effects, association effects and interaction effects) at the same time.

The basic idea of log-linear analysis is to search for the models that best fit the data. In order to do this, one needs to specify and compare all the models to each other. For this purpose, expected cell frequencies are generated for each model and the respective goodness-of-fit statistic is calculated. Two chi-square statistics can be used:

The familiar Pearson chi-square statistic
$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

and the Likelihood-ratio chi-square
$$L^2 = 2 \sum_i \sum_j O_{ij} \cdot \ln \frac{O_{ij}}{E_{ij}}$$

For large sample sizes these statistics are equivalent. The advantage of the likelihood-ratio chi-square however, is that it can be subdivided into interpretable parts that add up to the total. This property is very useful when comparing the different models.

Assumptions

Log-linear analysis requires no distributional assumptions. The only assumption needed is that the observations are independent.

Furthermore, there are two requirements that are easy to satisfy in the present study:

- All the cells in the contingency table should have expected frequencies greater than 1
- No more than 20% of the cells should have expected frequencies less than 5.

The procedure

The most common procedure to approach the best model is called Backward Elimination. In this procedure one starts with the most complex model (usually the saturated model which contains all the possible terms, including the main effects and all possible interactions between the variables) and eliminates effects from it one by one in a step-wise fashion. The comparison between successive models is done by subtracting the L^2 value of one from the L^2 value of the other and the degrees of freedom of the one from the degrees of freedom of the other. Then critical values from the chi-square distribution can be used to evaluate the significance of the residual L^2 from the residual degrees of freedom.

Another way to approach the best model is to test for the significance of the individual terms in the model. A partial chi-squares table is produced by SPSS indicating the significance of each main effect, association or interaction term in the model. From that table one can choose all the significant terms to make the best and most parsimonious model.

The researcher decided to use the second approach because:

- It is easier to understand
- It is easier to interpret the results of the analyses
- It is less time consuming.

4.1.10 Investigation of possible factors associated with misfit

Log-linear analysis was performed in an attempt to investigate possible associations of various factors with misfit in the maths test. Several models were used with different combinations of variables. The maximum number of variables used was 3 (except for one model) since higher level interactions are generally very difficult to interpret. The models considered included:

*Different schools * Different teachers * Misfit*

*Teacher gender * Student gender * Misfit*

*Student gender * Ability * Anxiety * Misfit*

*Student gender * ADHD * Misfit*

*Student gender * Study time * Misfit*

*Student gender * Private tuition * Misfit*

*Student * Item order * Misfit*

*Student gender * Atypical schooling * Misfit*

*Student gender * Is maths favourite * Misfit*

*Student gender * Revision periods before test * Misfit*

The combinations of variables were decided by the researcher in terms of the most likely combinations (in the researcher's opinion) to have an association with misfit.

Student gender was used in all models because:

- There were differences in the anxiety levels between genders with girls showing higher levels of anxiety ($p = 0.000$).
- There were differences in ability measures between genders, favouring the girls ($p = 0.012$).
- Girls demonstrated higher levels of language competency ($p = 0.000$).
- Girls spend more time studying for maths ($p = 0.013$).
- Higher proportion of girls takes private tuition in maths ($p = 0.040$).
- Higher proportion of girls shows interest in maths ($p = 0.005$)

All the models used (except from one) had 3 variables. The ability and anxiety variables were however combined together, and with gender and misfit, since there is a reference in the literature of an interaction effect of the two variables on misfit (Bracey and Rudner, 1992).

The results of the analyses are reported below (detailed tables can be found in the appendices).

Different schools * Different teachers * Misfit

No association between school and Misfit was found. However there was a significant association ($p = 0,027$) between teachers and misfit. Some teachers had higher proportions of misfitting students than others. No association between the interaction of schools and teachers on misfit were found.

Teacher gender * Student gender * Misfit

No significant association was found between teacher gender and misfit or student gender and misfit. The interaction of teacher and student gender on misfit was also non-significant.

Student gender * Ability * Test Anxiety * Misfit

No association was found between student gender, ability, test anxiety, or any combination of those variables, with misfit.

Significant association were found between gender and ability ($p = 0.000$) between gender and test anxiety ($p = 0.000$) and between ability and test anxiety ($p = 0.000$).

The results were very similar to the ones above both when the 20th and 80th percentiles and the 10th and 90th percentiles were used as cut-off scores for the ability and test anxiety groups.

Similarly, when the worry factor was isolated from the anxiety scale and used in the place of anxiety, the log-linear analysis showed very similar results with no association between gender, ability and worry or any combination of those variables with misfit.

Gender * ADHD * Misfit

The ADHD variable used was a dichotomous variable taking only the values 0 (no ADHD symptoms observed by the teachers) and 1 (ADHD symptoms).

No association between ADHD and misfit was found. Similarly the interaction of student gender and ADHD on misfit was non-significant.

The only significant association found was that of student gender and ADHD with boys' being observed by their maths teacher to display ADHD symptoms having almost double the proportion of girls' (40.7% vs 21.1%).

When the models

*Gender * Study time * Misfit*

*Student gender * Private tuition * Misfit*

*Student * Item order * Misfit*

*Student gender * Is maths favourite * Misfit*

*Student gender * Revision periods before test * Misfit*

*Student gender * Interest in maths * Misfit*

were considered, no association between any of the variables and misfit was found. Similarly the interaction of student gender and each variable on misfit was non-significant.

Do the same students misfit in different administrations of measurement instruments?

The next table, table 4.1.30 compares the percentages of fitting and misfitting students in the maths test and the TAI. The purpose of this comparison is to investigate whether the misfit is a consistent or inherent characteristic of some students.

The table shows that 34% of the fitting students in the TAI (135 out of 397) misfit in the maths test whereas 35.6% of the misfitting students in TAI (26 out of 73) misfit in the maths test.

Similarly, 15.2% of the fitting students in the maths test (47 out of 309) misfit in the TAI, whereas, 16.1% of the misfitting students in the maths test (26 out of 161) misfit in TAI.

Both of these results are not significant indicating no association between misfitting in the maths test and misfitting in the TAI.

Table 4.1.30 Maths Test Misfit * TAI Misfit Crosstabulation

		TAI Misfit		Total
		Fitting Students	Misfitting Students	
Maths Test Misfit	Fitting Students	262	47	309
	Misfitting Students	135	26	161
Total		397	73	470

Chi-square = 0.018, d.f. = 1, p = 0.895

Comparing the internal consistency of raw scores for fitting and misfitting students

Cronbach’s alpha was used as a measure of the internal consistency of the raw scores. At the same time, the standard error of alpha (ASE) and 95% confidence intervals for alpha were computed using the method suggested by Iacobucci and Duhachek (2003) in order to make comparisons possible.

First, alpha, ASE and confidence intervals were computed for two groups, the fitting and misfitting students. Table 4.1.31 shows the results. (N is the number of students in the group, K is the number of items, ASE is the standard error of alpha and low and high are the lower and higher limits of the 95% confidence intervals for alpha)

Table 4.1.31 95% C.I. for alpha for fitting and misfitting students

Student groups	Estimate of alpha	N	K	ASE	Low	High
Fitting	0.921	383	12	0.00596	0.909	0.933
Misfitting	0.880	189	12	0.0129	0.855	0.905

Inferences from comparisons of Confidence Intervals (CI)

Wainer (1996) describes various ways of depicting error for effective interpretations and correct inferences. In an example comparing maths scores from different states he constructs 95% confidence intervals of the states’ mean scores and states: “A difference between two states is statistically significant (at the 0.05 level) if one state’s data point is outside the other state’s bounds” (p. 107).

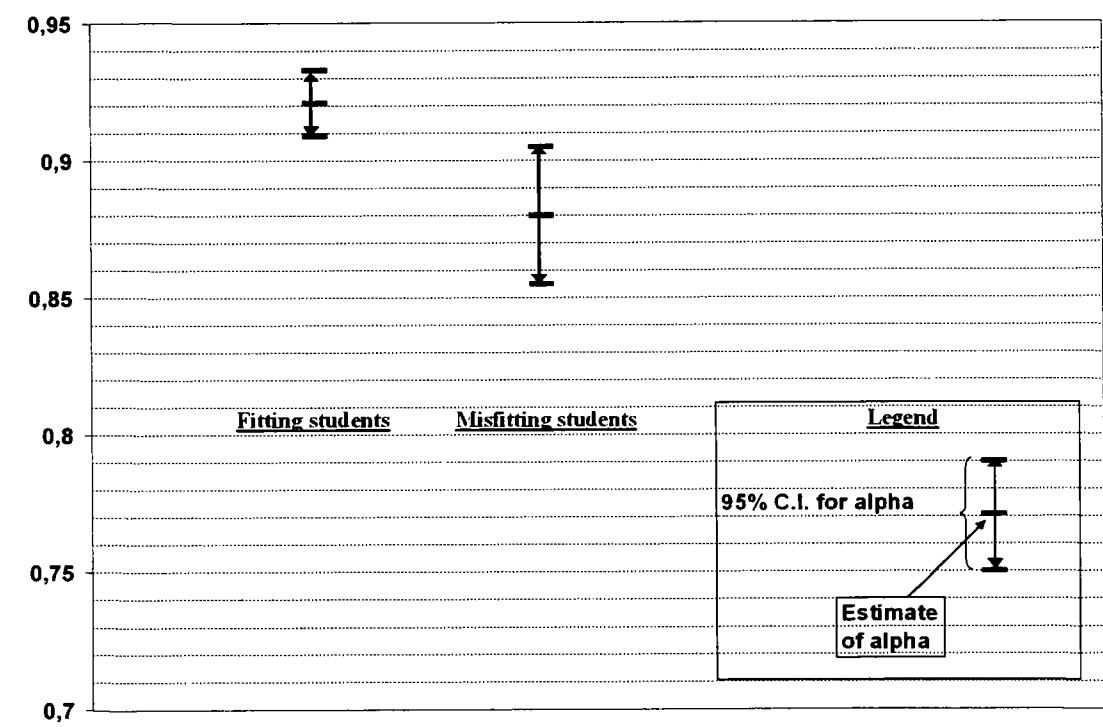
In the case where the confidence intervals are constructed by adding and subtracting 1 standard error he explains that for significant differences one should look for non-overlapping error bars.

In this study the hypothesis is that the internal consistency of misfitting students is of a lower degree than that of fitting students. Therefore, a 95% confidence interval for alpha of fitting students is constructed and if the alpha estimate of misfitting students (and later the alpha estimates of 3 groups of misfitting students: one with high outfit

values only, one with high infit values only and one with both high) is below the bounds of the 95% CI of the fitting students then there are significant differences with alpha being lower for the misfitting students.

Figure 4.1.16 shows the confidence intervals in a diagrammatic form. It is clear that the alpha estimate for the group of misfitting students is below the lower limit of the C.I. of alpha for the group of fitting students. Therefore we can conclude that there are differences between the alphas for the two groups with the alpha estimate of the misfitting students being significantly lower.

Figure 4.1.16 95% C.I. for alpha for fitting and misfitting students



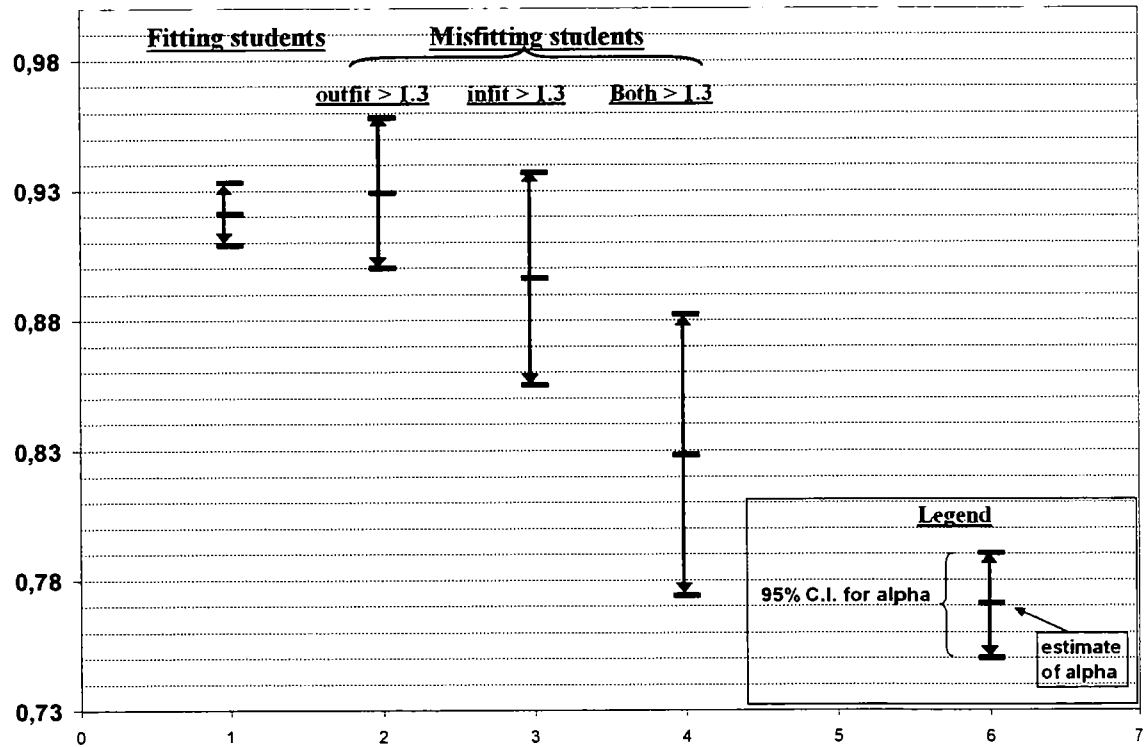
Second, alpha, ASE and confidence intervals were computed for four groups. The first was, as before, the fitting students but then the misfitting students group was divided into three groups. The one was students misfitting because of large outfit values, the second because of large infit values and the third because of a combination of large infit and outfit values. Table 4.1.32 below shows the results.

Table 4.1.32 95% C.I. for four groups of students.

Student groups	Estimate of alpha	N	K	ASE	Low	High
Fitting	0.921	383	12	0.00596	0.909	0.933
Misfitting (large outfit)	0.929	50	12	0.01483	0.900	0.958
Misfitting (large infit)	0.896	53	12	0.0210	0.855	0.937
Misfitting (large infit and outfit)	0.828	86	12	0.0274	0.774	0.882

Figure 4.1.17 shows the confidence intervals in a diagrammatic form. No significant differences are evident between the alphas of the fitting students and the first group of misfitting students (the ones with outfit > 1.3, i.e. the ones who misfit because of high outfit value only). However, misfit because of large infit values seems to be producing a large decrease in the internal consistency of the raw scores of the students, causing significant differences from the fitting students. This significant decrease is evident if infit is large, and even larger if both infit and outfit are large.

Figure4.1.17 95% C.I. for alpha for four groups of students



This is indicative of the effect of the two mean squares on the reliability of the response patterns of misfitting students in a short classroom test (12 multistep maths items).

The reliability of misfitting students with large infit values is significantly lower than fitting students, whereas large outfit values, on their own do not seem to affect the reliability of the students' responses.

Comparing alphas with simulated data

In order to investigate further the effect of infit on the reliability a dataset with 12 items and 2000 students was simulated, using WINSTEPS.

For comparison purposes, figures 4.1.18 and 4.1.19 show two item-person maps, one for the test data and one for the simulated data.

The items were anchored from the test calibrations and are therefore the same. Although 2000 students were used in the simulated data, it is clear that the distributions of abilities of the two datasets are very similar. Therefore the Rasch-fitting simulated data consisted of the 12 test items (they were anchored from the test calibration) and 2000 students with the same ability distribution as the students who took the test.

Misfitting students in the simulated data were identified in exactly the same way as in the test (with infit and outfit mean square values greater than 1.3).

Figure 4.1.18 *Test 1: Item-student map* Figure 4.1.19 *Simulated data: Item-stud. map*

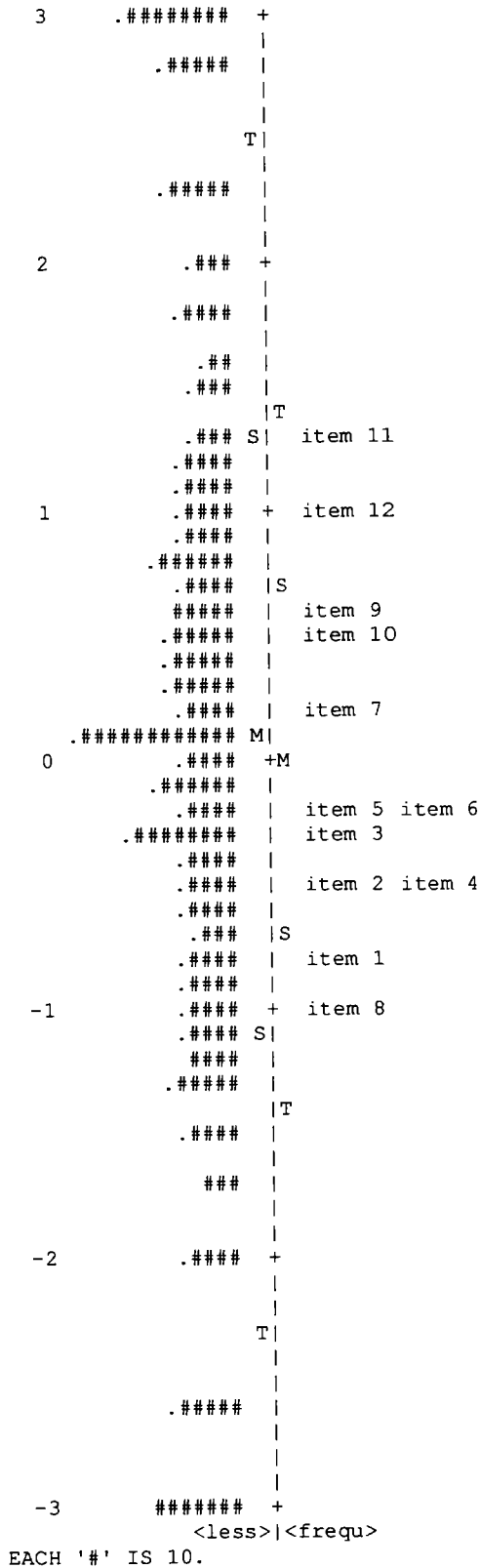
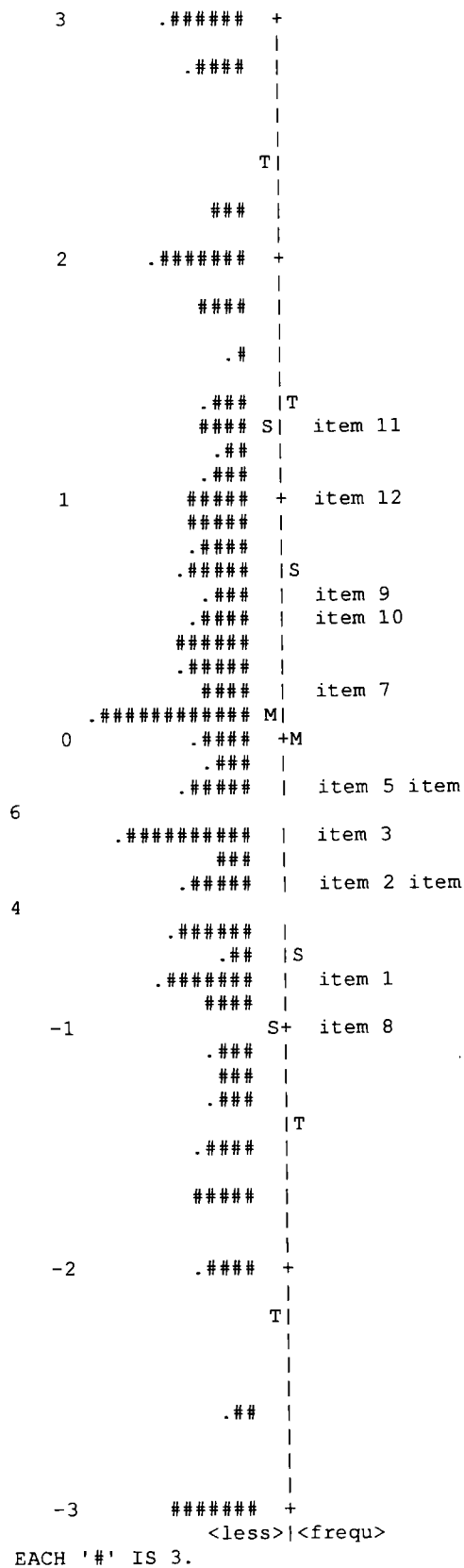


Table 4.1.33 shows the alpha estimates, ASE and the 95% confidence interval for alpha for two groups of students, the fitting and the misfitting students. The alpha estimates are almost identical with the test data.

Table 4.1.33 95% C.I. for alpha for fitting and misfitting students in the simulated data

Student groups	Estimate of alpha	N	K	ASE	Low	High
Fitting	0.928	1536	12	0.00271	0.923	0.933
Misfitting	0.878	464	12	0.00837	0.862	0.894

Table 4.1.34 is similar to table 4.1.33 but for four groups of students, the fitting and the misfitting students because of large outfit, infit and both outfit and infit values. The alpha estimates are almost identical with the test data also for the group of students who misfit because of large outfit values only. The alpha estimate is now slightly lower for the misfitting (by large infit) students (0.884 compared with 0.896) and much lower for the misfitting (by large outfit and infit) students (0.778 compared with 0.828)

Table 4.1.34 95% C.I. for four groups of students in the simulated data

Student groups	Estimate of alpha	N	K	ASE	Low	High
Fitting	0.921	1536	12	0.00271	0.923	0.933
Misfitting (large outfit)	0.934	144	12	0.00812	0.918	0.950
Misfitting (large infit)	0.884	145	12	0.0142	0.856	0.912
Misfitting (large infit and outfit)	0.778	175	12	0.0248	0.729	0.827

Figures 4.1.20 and 4.1.21 show the alpha estimates and the 95% confidence intervals in the 2 cases.

Figure 4.1.20 95% C.I. for alpha for fitting and misfitting students (simulated data)

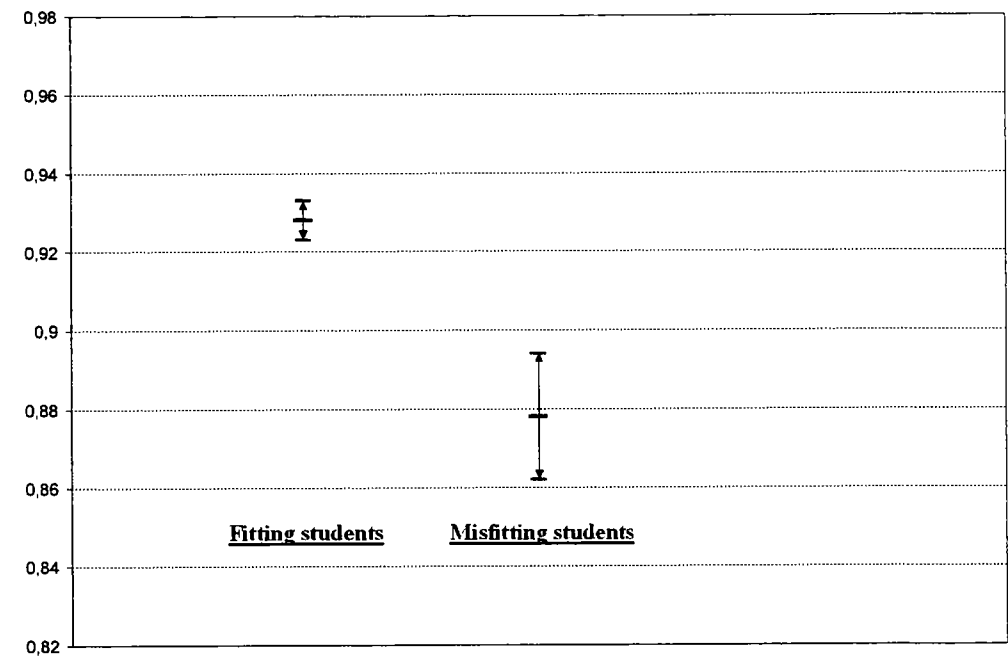


Figure 4.1.21 95% C.I. for alpha for four groups of students (simulated data)

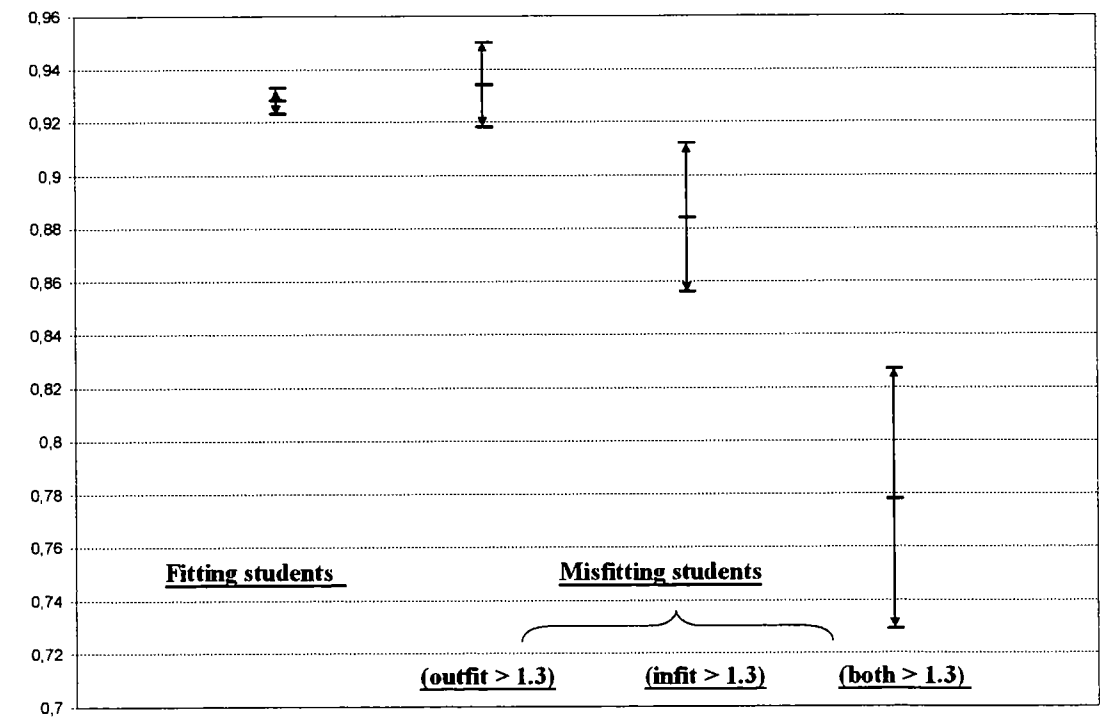


Figure 4.1.20 shows the alpha confidence intervals in a diagrammatic form. It is clear that the alpha estimate for the group of misfitting students is significantly lower (since is below the lower limit of the C.I. of alpha for the group of fitting students).

In figure 4.1.21 no significant differences are evident between the alphas of the fitting students and the first group of misfitting students (the ones with large outfit values only). However, misfit because of large infit values seems to be producing a large decrease in the internal consistency of the raw scores of the students, causing significant differences from the fitting students. This significant decrease is evident if infit is large, and even larger if both infit and outfit are large.

The results from the simulated data investigation are in agreement with the results from the investigation on the test data. Both analyses show that the reliability of the scores of the misfitting students is lower than that of the fitting students and that it seems to be affected mostly by large infit values.

4.2 PHASE 2 RESULTS

4.2.1 The First Maths Test (The Diagnostic Test) and its calibration

The sample

The test was administered to 635 students in 3 schools: Limassol 1 (5 teachers, 12 classes and 306 students), Limassol 2 (1 teacher, 2 classes and 23 students) and Paphos (7 teachers, 11 classes and 286 students). A total of 13 teachers and 25 classes were involved. Overall, out of the total of 635 students, 43.9% were male and 56.1% female.

Table 4.2.1 shows the distribution of the 635 students by gender, in the 3 different schools. In all schools the proportion of female students was larger than that of male students.

Table 4.2.1 School * Gender Crosstabulation

		School			Total
		Limassol 1	Limassol 2	Paphos	
Gender	Male	122	20	137	279
	Female	184	23	149	356
Total		306	43	286	635

Test calibrations

The Rasch model was used for the calibrations. The first calibration on the full dataset revealed two misfitting items ($1.5 < \text{outfit} < 1.98$), 3 marginally misfitting items ($1.3 < \text{outfit} < 1.5$) and 19 badly misfitting students (outfit and/or $\text{infit} > 2.7$).

The 19 students were removed and a second calibration was performed, revealing only 3 slightly misfitting items. Those items were retained in the dataset (the reasons for not removing the items are explained).

The item statistics from the second calibration were then used for the final calibration in order to get the students statistics.

Item-person maps are presented to show how well the items were targeted for the population of students and finally the students were divided into groups according to their ability for investigating later on whether ability is associated with misfit.

First calibration

The first calibration, in which the full set of the test data was used (27 items and 635 students), revealed two misfitting items, items 10 and 2a ($\text{outfit} > 1.5$) and 3 slightly misfitting items, items 7, 2c and 9a, ($1.3 < \text{outfit} < 1.5$) as shown in table 4.2.2.

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.

Table 4.2.2 Items Statistics: Misfit order

ENTRY	RAW					INFIT	OUTFIT	PTMEA		
NUMBER	SCORE	COUNT	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	ZSTD	CORR.	items
27	2055	618	.47	.04	1.50	6.2	1.95	5.2	A .67	item 10
6	554	618	-1.67	.14	1.14	1.4	1.74	3.0	B .27	item 2a
23	1903	618	-.16	.05	1.36	4.5	1.49	4.1	C .63	item 7
8	231	618	1.68	.09	1.07	1.8	1.41	5.0	D .39	item 2c
25	522	618	1.34	.06	1.16	2.9	1.35	2.4	E .56	item 9a
7	557	618	-1.74	.15	1.13	1.3	1.29	1.3	F .29	item 2b
24	1018	618	.90	.04	1.18	3.0	1.20	1.6	G .65	item 8
1	566	618	-1.94	.16	1.08	.8	1.16	.7	H .32	item 1a
20	300	618	1.11	.09	1.06	1.7	1.07	1.2	I .43	item 5c
26	1381	618	1.25	.03	1.05	.8	1.03	.4	J .74	item 9b
11	488	618	-.65	.11	1.04	.7	.99	.0	K .41	item 3c
15	1487	618	.65	.04	1.00	.0	1.01	.2	L .71	item 3g
18	318	618	.96	.09	.95	-1.3	.93	-1.2	M .50	item 5a
9	571	618	-2.07	.16	.95	-.4	.61	-1.7	N .41	item 3a
19	334	618	.83	.09	.95	-1.5	.95	-.9	m .50	item 5b
21	326	618	.89	.09	.95	-1.5	.88	-2.1	l .51	item 5d
13	848	618	.25	.06	.93	-1.2	.84	-1.4	k .63	item 3e
3	528	618	-1.20	.13	.92	-1.0	.92	-.4	j .45	item 1c
4	439	618	-.11	.10	.92	-1.6	.92	-.9	i .50	item 1d
12	385	618	.39	.09	.92	-2.0	.91	-1.3	h .51	item 3d
14	713	618	.74	.06	.90	-2.1	.86	-1.4	g .64	item 3f
22	859	618	.21	.06	.89	-1.9	.82	-1.5	f .63	item 6
2	564	618	-1.90	.15	.89	-1.0	.64	-1.7	e .44	item 1b
10	571	618	-2.07	.16	.89	-.9	.50	-2.4	d .45	item 3b
5	378	618	.45	.09	.85	-4.0	.77	-3.8	c .56	item 1e
17	1195	618	.53	.05	.82	-2.9	.72	-1.9	b .70	item 4b
16	667	618	.88	.06	.77	-5.1	.69	-4.0	a .68	item 4a
MEAN	732.	618.	.00	.09	1.01	-.1	1.02	-.1		
S.D.	472.	0.	1.17	.04	.16	2.5	.33	2.4		

Table 4.2.3 shows the top part of the table with the student statistics from the original calibration in misfit order (outfit > 2.3 and/or infit > 2.3). The first 19 students in table 4.2.3 with outfit (or infit) > 2.7 (3.0%) were considered badly misfitting and distorting the calibration process.

Table 4.2.3 Student statistics: misfit order

ENTRY	RAW					INFIT		OUTFIT	PTMEA	
NUMBER	SCORE	COUNT	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	ZSTD	CORR.	stude
189	4	27	-2.20	.59	3.60	3.7	7.79	3.2	A .24	10810
379	49	27	3.61	.99	1.03	.4	7.31	2.2	B-.28	30202
622	48	27	2.94	.69	.82	.1	5.04	2.1	C-.28	31111
138	48	27	2.94	.69	.99	.3	4.29	1.9	D-.15	10609
335	11	27	-.68	.36	3.41	3.2	2.27	2.3	E .41	20608
21	42	27	1.68	.34	3.27	3.1	2.49	1.9	F .01	10122
369	4	27	-2.20	.59	3.27	3.4	1.62	.9	G .28	30120
390	38	27	1.30	.28	1.74	1.6	3.21	3.0	H-.03	30213
374	47	27	2.57	.55	1.43	.8	3.20	1.7	I-.05	30125
139	46	27	2.30	.48	.88	.1	3.13	1.8	J-.11	10610
125	11	27	-.68	.36	1.86	1.6	3.09	3.3	K .09	10522
483	47	27	2.57	.55	.87	.1	3.00	1.6	L .04	30527
255	45	27	2.10	.42	1.25	.6	2.94	1.9	M-.02	11025
222	45	27	2.10	.42	.86	.0	2.93	1.9	N-.14	10918
531	3	27	-2.59	.66	1.02	.2	2.91	1.5	O .23	30719
447	3	27	-2.59	.66	1.01	.2	2.91	1.5	P .23	30413
627	39	27	1.39	.29	2.88	3.1	1.39	.8	Q .34	31117
176	40	27	1.47	.31	.88	-.1	2.78	2.3	R .11	10722
147	9	27	-.98	.41	2.08	1.8	2.73	2.4	S .41	10618
413	48	27	2.94	.69	.94	.2	2.68	1.3	T-.03	30307
615	41	27	1.57	.32	2.65	2.6	1.48	.9	U .33	31101
166	7	27	-1.37	.47	1.64	1.2	2.62	1.9	V .39	10711
562	33	27	.95	.25	.89	-.2	2.60	2.9	W .11	30825
286	46	27	2.30	.48	2.25	1.6	2.60	1.5	X .11	11205
380	43	27	1.80	.36	1.90	1.5	2.55	1.8	Y .00	30203
41	44	27	1.94	.39	2.53	2.1	.57	-.4	Z .29	10215
106	14	27	-.34	.32	2.43	2.5	.96	.0	.84	10503
27	10	27	-.82	.39	2.41	2.2	1.57	1.2	.31	10201
633	40	27	1.47	.31	2.41	2.4	1.43	.9	.31	31124
344	36	27	1.15	.27	.92	-.1	2.36	2.3	.01	20617
465	13	27	-.44	.33	1.69	1.4	2.34	2.7	.51	30504
536	32	27	.88	.25	2.30	3.1	1.06	.3	.56	30725

These 19 students were therefore removed, leading to a second calibration with again the 27 items, but this time with 616 students.

Second calibration

Table 4.2.4 below shows the item statistics from this second calibration. This time, only three items were slightly misfitting, items 10, 7 and 9a, with the outfit statistics in the range 1.3 – 1.52 and the infit statistics being 1.43 and 1.37 for the first two items. Item 9a (outfit = 1.32) was only marginally misfitting.

Table 4.2.4 Item statistics: misfit order

ENTRY	RAW					INFIT		OUTFIT	PTMEA		
NUMBER	SCORE	COUNT	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	ZSTD	CORR.	items	
27	1998	599	.48	.04	1.43	5.4	1.52	3.1	A .68	item 10	
23	1851	599	-.14	.05	1.37	4.6	1.48	4.0	B .61	item 7	
25	498	599	1.39	.06	1.16	2.9	1.32	2.3	C .55	item 9a	
7	544	599	-1.78	.15	1.17	1.5	1.27	1.2	D .26	item 2b	
24	985	599	.92	.05	1.20	3.2	1.24	2.0	E .64	item 8	
8	218	599	1.75	.09	1.07	1.7	1.21	2.8	F .40	item 2a	
6	545	599	-1.80	.15	1.15	1.4	1.20	.9	G .27	item 2a	
20	292	599	1.12	.09	1.07	1.8	1.09	1.5	H .42	item 5c	
11	476	599	-.65	.11	1.06	.9	1.01	.2	I .39	item 3c	
26	1323	599	1.29	.04	1.05	.9	.97	-.3	J .74	item 9b	
1	551	599	-1.95	.16	1.04	.4	.77	-1.0	K .35	item 1a	
15	1444	599	.67	.04	1.00	.1	1.03	.4	L .71	item 3g	
19	322	599	.86	.09	.96	-1.0	.97	-.5	M .48	item 5b	
9	559	599	-2.18	.17	.96	-.3	.55	-1.9	N .40	item 3a	
21	313	599	.94	.09	.96	-1.3	.90	-1.9	m .50	item 5d	
18	309	599	.97	.09	.95	-1.4	.93	-1.2	l .50	item 5a	
13	826	599	.26	.06	.94	-1.0	.86	-1.2	k .61	item 3e	
4	427	599	-.10	.10	.93	-1.4	.94	-.7	j .49	item 1d	
3	516	599	-1.23	.13	.94	-.7	.93	-.4	i .43	item 1c	
12	375	599	.40	.10	.92	-2.1	.90	-1.5	h .51	item 3d	
22	835	599	.23	.06	.91	-1.6	.85	-1.3	g .62	item 6	
2	550	599	-1.93	.16	.90	-.8	.67	-1.5	f .42	item 1b	
10	559	599	-2.18	.17	.90	-.7	.47	-2.3	e .43	item 3b	
14	687	599	.77	.06	.90	-2.0	.86	-1.4	d .63	item 3f	
5	369	599	.45	.09	.85	-3.9	.78	-3.6	c .56	item 1e	
17	1159	599	.55	.05	.83	-2.7	.75	-1.8	b .69	item 4b	
16	648	599	.89	.06	.77	-5.0	.70	-3.9	a .68	item 4a	
MEAN	710.	599.	.00	.09	1.01	-.1	.97	-.3			
S. D.	458.	0.	1.21	.04	.15	2.3	.25	1.9			

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 although it was an item on a rather specific content it was decided not to 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.

Overall it was an item of just above average difficulty (difficulty 0.30) and 52% of the students scored full marks whereas 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 a rather easier item (difficulty of -0.33) on simple geometry; 57% of the students scored full marks, whereas only 8% scored zero marks. This item required knowledge of basic angle properties like:

- sum of angles in a triangle
- vertically opposite angles
- angles on a line and
- angles on parallel lines which are cut by a transversal.

Misfit has probably 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 the fact that if one or more terms in an equation are algebraic fractions, then the roots of the denominators are values of x that cannot be in the possible solution range. The difficulty of this item is also evident in the proportion of students scoring full marks. Only 38% managed to score the full marks (2 marks) and 8% gave a half-correct answer scoring one mark. The remaining 53% of the students scored no marks. Misfit in this item most probably occurred because of a few unexpected correct answers by students copying from more knowledgeable classmates.

The three above-mentioned 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 2nd calibration is given in table 4.2.6.

Table 4.2.6 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)
Examinees	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)

The range of student abilities was from -2.63 to 3.64, with a mean of 1.03 (SD = 1.17). The reliability of student estimates 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 4 statistically distinct strata ($\text{strata} = 3.81$) 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 the 635 students.

Third and final calibration

Figure 4.2.1 shows the item-student map. One can see that most of the test items are well targeted for students with abilities around and below the mean ability. Only 4 out of the 27 items are targeted for 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 for students with ability of more than 2 standard deviations below the mean ability.

The wide targeting of the items is perhaps easier to see in figure 4.2.2 which shows another item – student map, this time with all the categories of the items (the thresholds for all the possible scores for each item). It is obvious that the various steps of the questions are well targeted for a wider range of abilities, from 3 standard deviations below to just above one standard deviation above the overall mean ability.

Given the purpose of the test (the test was a diagnostic test aiming mainly to discover the weakest students and their weaknesses, in order to be able to help them through extra lessons provided by the school), the targeting of the items is very satisfactory.

Figure 4.2.1 Students map of items

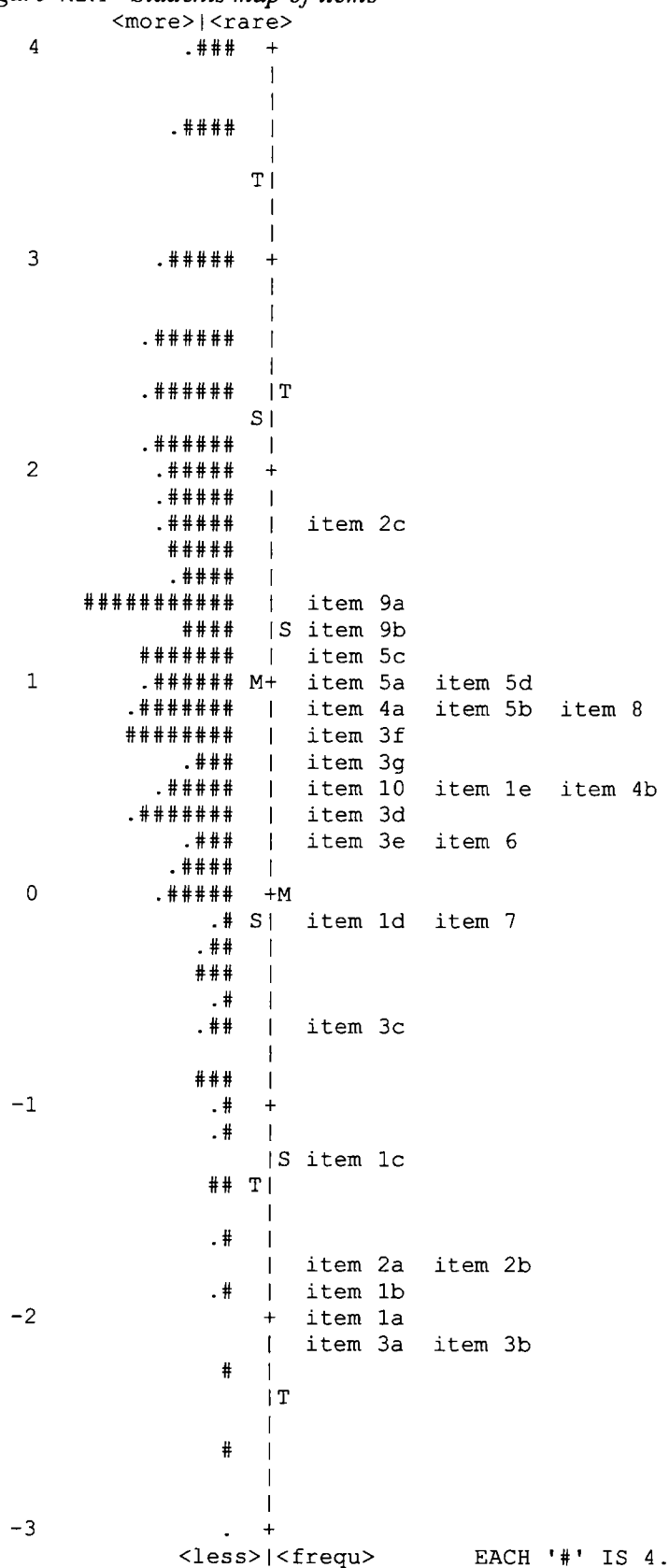
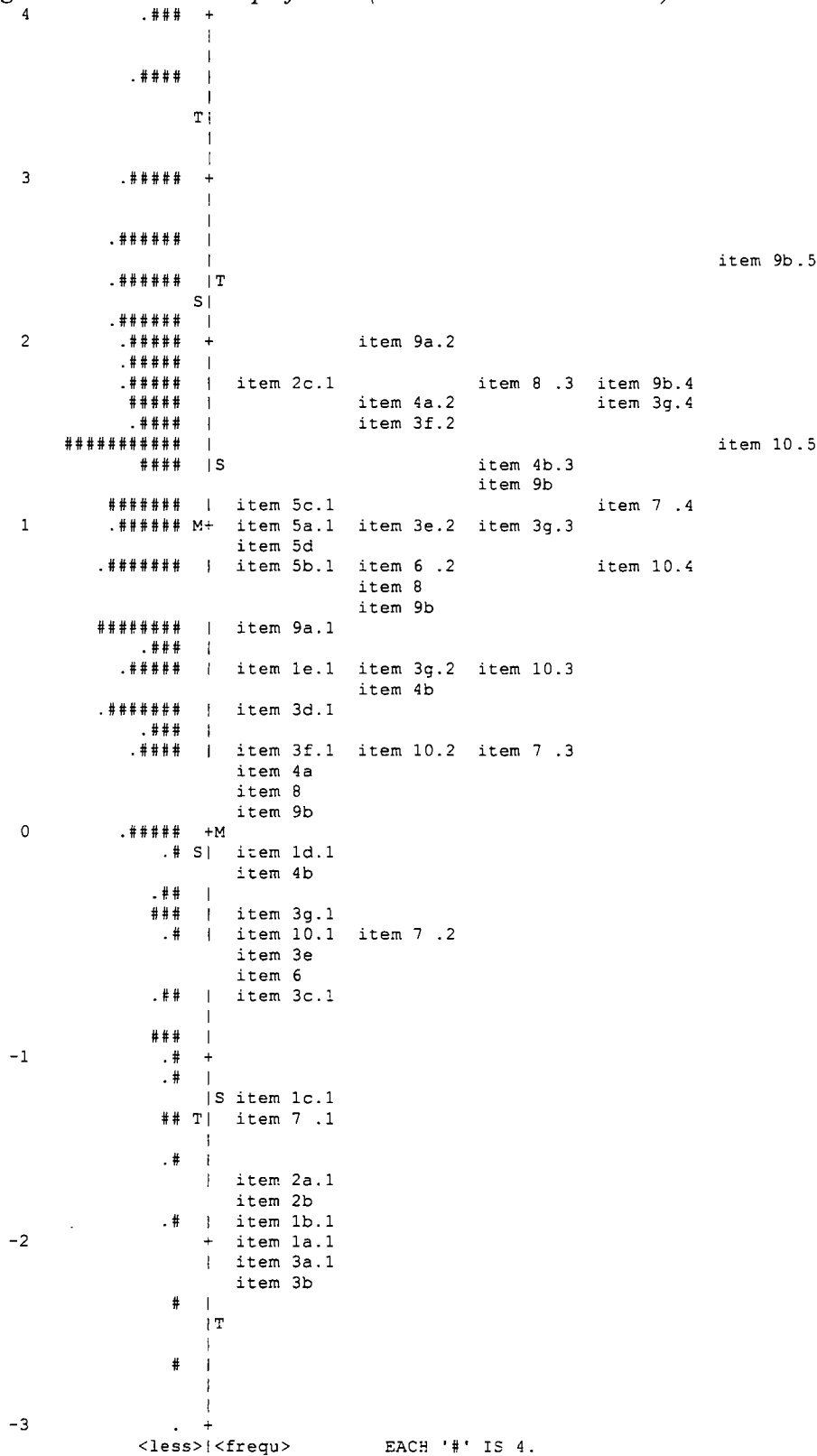


Figure 4.2.2 students map of items (with item score thresholds)



4.2.2 Reliability and validity of the first test

For the study of the reliability of the test two equivalent indices were used: the student reliability (this index is given by the Rasch analyses) and Cronbach's alpha.

For the validation study of the test, the following evidence was collected and presented below:

- *Principal components analysis of the standardized residuals after the Rasch calibrations, as proposed by Linacre (1998a).*
- *A plot of the factor loadings (on the first dimension extracted, other than the dimension measured by the test) against item measures.*
- *Correlations of the maths test scores with the final maths exam scores.*
- *Comparisons of the item estimates from two different calibrations (using two different samples based on the students' gender) to ascertain whether invariance holds.*

Reliability of the test

The student reliability was 0.87, as shown in table 4.2.6. This index is an indication of the precision of the instrument and shows how well the instrument can distinguish individuals.

Cronbach's alpha was high (0.901) indicating also a high degree of reliability (such alpha is acceptable even for high stakes tests). Alpha is a measure of the internal consistency of the test.

Validity of the test

Principal components analysis on the standardised residuals (Linacre, 1988) was performed in WINSTEPS yielding:

PRINCIPAL COMPONENTS (STANDARDIZED RESIDUAL) FACTOR PLOT				
Factor 1 extracts 2.0 units out of 27 units of item residual variance noise.				
Yardstick (variance explained by measures)-to-This Factor ratio: 57.9:1				
Yardstick-to-Total Noise ratio (total variance of residuals): 4.3:1				
Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)				
		Empirical		Modeled
Total variance in observations	=	142.5	100.0%	100.0%
Variance explained by measures	=	115.5	81.0%	80.5%
Unexplained variance (total)	=	27.0	19.0%	19.5%
Unexpl var explained by 1st factor	=	2.0	1.4%	

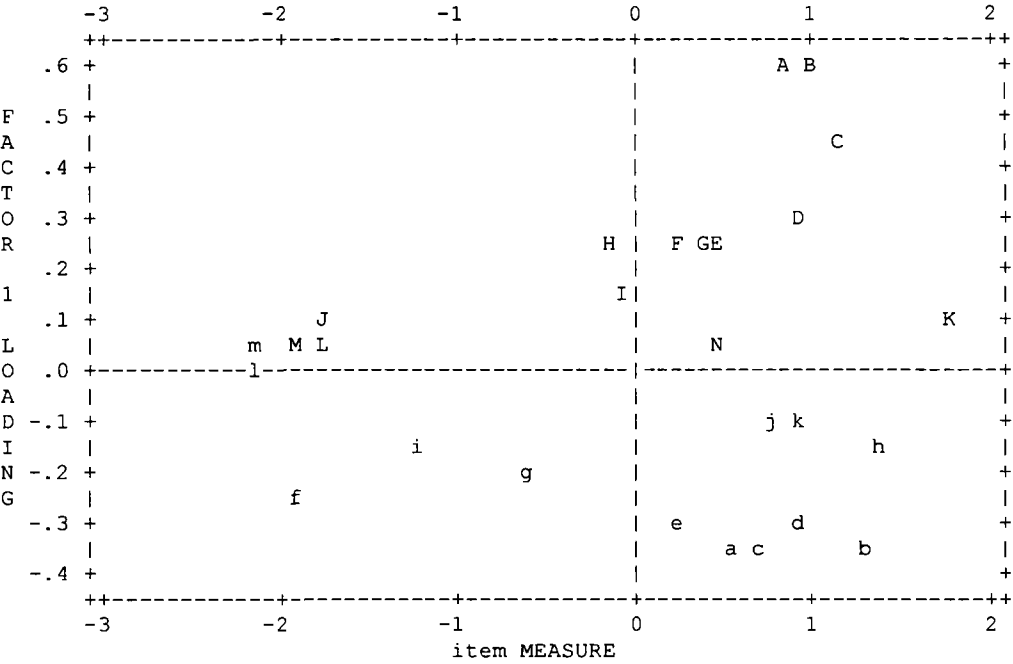
The variance explained by the measures (i.e. by the dimension measured by the test) is 81.0% of the total variance. It is also about 58 times the variance explained by the first factor extracted by PCA on the standardised residuals and about 4 and a half times the total unexplained variance in the data. The unexplained variance is 19% of the total variance in the data.

Also, the variance explained by this first factor is 7.4% of the unexplained variance (2 out of 27) and just 1.4% of the total variance in the data.

All of the above support the hypothesis that there is no second dimension present in the data, therefore the test is unidimensional.

This hypothesis of a unidimensional assessment is supported further by figure 4.2.3 below which shows the plot of the items loadings on the first factor extracted against the items' measures.

Figure 4.2.3 Factor 1 loadings against item measures.



One can safely say that there are no obvious item groupings, with items loading significantly on this first factor, in this plot.

Furthermore, the scores on the test were compared with the final mathematics exam results of the students in the 3 schools. This was done separately for each school since each school prepared its own final examination. The correlation coefficients (all highly significant) were:

Limassol 1: $r = 0.795$ ($N = 287$), Limassol 2: $r = 0.704$ ($N = 37$), Paphos: $r = 0.701$ ($N = 281$)

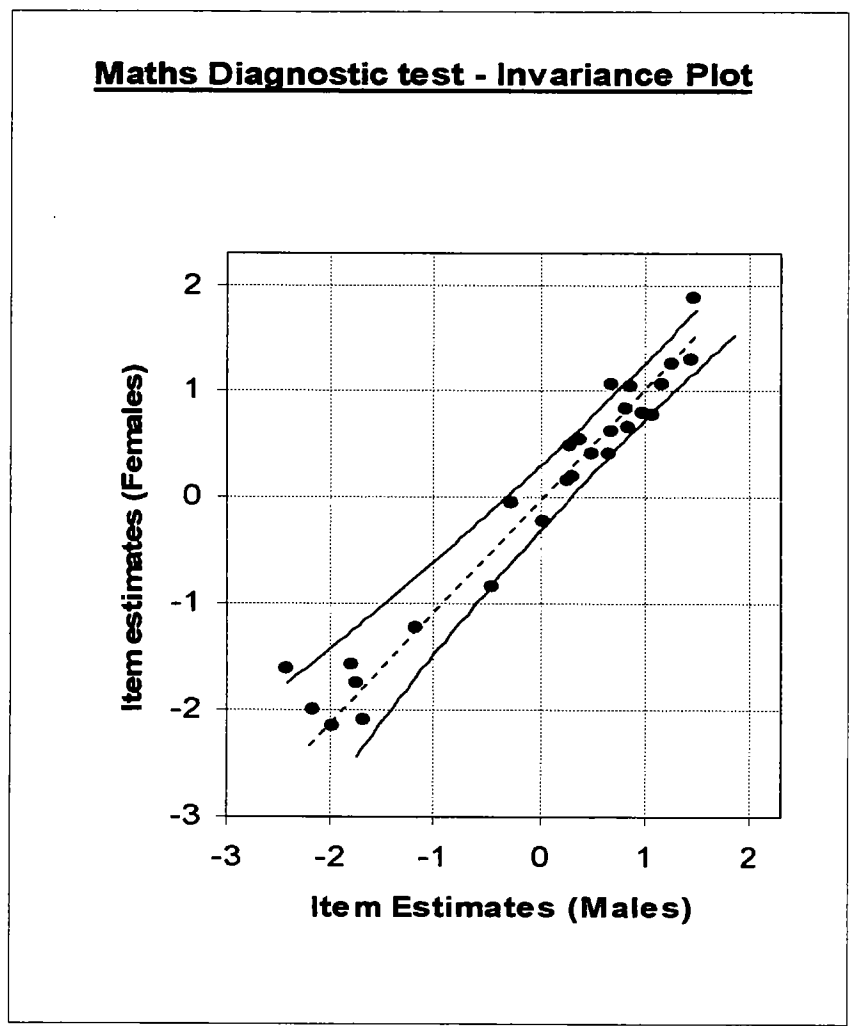
One would perhaps expect even higher correlations between the two maths test scores, however the first test (the diagnostic test) was a very easy test, targeted for the lower ability students, whereas the final maths exam was not. Therefore some lower-ability students who performed relatively well in the first test did not do so in the final exam, thus lowering a little the correlation.

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, because of too many unauthorised absences.

Comparisons of item estimates from two calibrations
Split of the data by gender

In this case the data was split into two groups based on gender. The two groups had sizes 279 (males) and 356 (females). Figure 4.2.5 below shows the invariance plot for the item estimates from these two subsets.

Figure 4.2.5 Invariance plot for the diagnostic maths test(by gender)



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 confidence limits, and that is a good indication that invariance holds. Also, the correlation coefficient is 0.975 which is extremely high.

Testing whether 3 items (out of 27) outside the 95% C.I. is unexpected ($p < 0.05$)

In a binomial situation where one has 27 items, each with $P(\text{lying outside C.I.}) = 0.05$. The expected number of items lying outside the C.I. is 1.35.

Let X = number of items outside the 95% C.I.

$H_0: p = 0.05$ (Under $H_0: X \sim \text{Bin}(27, 0.05)$)

$H_1: p > 0.05$

$P(X \geq 3) = 0.15 \gg 0.05$, therefore we cannot reject H_0 .

Conclusion: Three points outside the 95% C.I. is not a highly unlikely event if one has 27 items.

These results support the property of invariance of the Rasch model. This invariance of item calibrations across groups implies that the construct measured by the instrument has the same meaning to the groups studied.

All of the above evidence, together with the fact that the diagnostic test was constructed in accordance to the guidelines from the ministry of education and with the cooperation of the maths teachers in the three schools and the good fit of the test data to the Rasch model, support the belief of a high degree of validity.

Ability groups

The range of abilities was divided into three different groups, the low, medium and top ability groups using three different cut-off scores.

First, the range of abilities was divided into 3 groups using the 10th (measure of -0.678) and 90th (measure of 2.592) percentiles. The lowest 10% of the distribution was labelled the 'Low 10% Ability' group, the middle 80% the 'Medium Ability' group and the top 10% the 'Top 10% Ability' group.

Second, the range of abilities was divided into 3 groups using the 20th (measure of 0.093) and 80th (measure of 2.129) percentiles.

Finally, the range of abilities was divided into 3 groups using the 30th (measure of 0.482) and 70th (measure of 1.705) percentiles.

In both of the last two cases, again the groups were labelled Low, Medium and Top ability.

4.2.3 Misfitting students

Misfitting students were identified using appropriate cut-off scores for the infit and outfit statistics (1.3 for both). The numbers and proportions of misfitting students are presented, together with comparisons of equivalent proportions from a simulation study.

Following the calibration of the test, misfitting students were identified using cut-off scores for the infit and outfit mean squares of 1.3.

Table 4.2.8 shows the number of students identified as misfitting by the two indices as well as the total number.

Table 4.2.8 Misfit (infit) * Misfit (outfit) Crosstabulation

		Misfit (outfit)		Total
		Fitting	Misfitting	
Misfit (infit)	Fitting	413	74	487
	Misfitting	77	71	148
Total		490	145	635

The number of students identified as misfitting by the infit statistic was 148 (23.3%) and by the outfit statistic was 145 (22.8%) whereas 71 students were identified by both, giving a total of 222 (34.9%) misfitting students.

A simulation study was carried out. WINSTEPS (Linacre, 2005) provides users the opportunity to use the estimated person, item and structure measures to simulate a Rasch-fitting data set equivalent to the raw data. This can be used to investigate the stability of measures and distribution of fit statistics.

The infit mean square calculated for this Rasch-fitting data set identified 13.7% misfitting students (infit > 1.3) and the outfit mean square 20.2 % (outfit > 1.3). The proportion of misfitting students in the simulated dataset was 28.2%.

The infit and outfit mean square statistics follow a Chi-square distribution with expected value of 1. Even when the data fit the Rasch model perfectly, whatever cut-off scores for identifying misfitting examinees are used (1.2, 1.3, 1.4 or higher) there will be a proportion of examinees with mean square statistics greater than the cut-off score, thus labelled misfitting. The higher the cut-off scores the lower the proportion of misfitting students. In other words, whatever cut-off score is used the Rasch model expects a proportion of examinees to have aberrant responses

The results of the simulation study show a similar proportion for the outfit and a lower proportion for the infit than the results from the analyses of the test data. The overall proportion of misfitting students in the simulated data is slightly lower than that in the actual test data. The two proportions (for infit and overall) were lower than the proportions found in the empirical data since simulated data are always expected to fit the Rasch model better.

The simulation study shows that although the data fit the Rasch model satisfactorily, the fit is, as most probably expected for empirical data, not perfect.

4.2.4 Maths Self-esteem scale (MSES)

The sample

The MSES consisted of 6 items and was administered to 553 students out of the 635 who took the diagnostic test. In Limassol 1, 277 students completed the scale, in Limassol 2, 39 and in Paphos 237.

Table 4.2.9 shows the number of male and female students who were administered the scale, in the 3 different schools. The proportions of male and female students in this sample are 42.7% and 57.3% respectively. These proportions are similar to the proportions of the original sample taking the diagnostic test (43.9% and 56.1%).

Table 4.2.9 School * Gender Crosstabulation

		Schools			Total
		Limassol 1	Limassol 2	Paphos	
Gender	Male	109	17	110	236
	Female	168	22	127	317
Total		277	39	237	553

There were two reasons why the sample of students answering the MSES was by 82 smaller than the original sample. Those were:

- One teacher who taught one class in the Paphos school did not want to administer the MSES to her 27 students and
- 55 students were either absent when the MSES was administered or did not want to complete it.

MSES calibrations

The Rasch RSM was used for the calibrations. The first calibration on the full dataset revealed no misfitting items therefore no other calibration was considered necessary.

Item-person maps are presented to show how well the items are targeted for the population of students and finally the students are divided into groups according to their maths self-esteem estimates for investigating later on whether maths self-esteem is associated with misfit.

The first calibration of the full set of data included 548 students (5 students, 2 maximum scorers and 3 minimum scorers were removed from the calibration) and 6 items.

Table 4.2.10 below shows the item statistics in misfit order.

Table 4.2.10 Items statistics: misfit order

ENTRY	RAW					INFIT	OUTFIT	PTMEA	
NUMBER	SCORE	COUNT	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	ZSTD	CORR.
6	2314	548	-.48	.05	1.34	5.1	1.32	4.6	A .69
5	2212	548	-.26	.05	1.18	2.9	1.13	2.0	B .75
4	1415	548	1.38	.05	.98	-.3	1.03	.4	C .74
2	1835	548	.51	.04	.90	-1.7	.96	-.6	c .78
3	2489	548	-.89	.05	.95	-.8	.88	-1.8	b .76
1	2208	548	-.25	.05	.68	-5.9	.69	-5.6	a .82
MEAN	2079.	548.	.00	.05	1.01	-.1	1.00	-.2	
S.D.	355.	0.	.74	.00	.21	3.5	.20	3.2	

There are no misfitting items; therefore the scale data fit the Rasch model very well. (Since this is a questionnaire and not a test, the same infit and outfit values as with the TAI, of 1.5, were used as the cut-off scores for identifying misfit). A summary of the results of the Rasch analysis from this calibration is given in table 4.2.11

Table 4.2.11 Summary of the results of the Rasch analysis for the MSES

	N	Estimate of mean (SD)	Range	Reliab.	Separ. Index	Infit msq mean (SD)	Outfit msq mean (SD)
Examinees	553	0.29 (1.39)	-3,62 to 3.89	0.83	2.25	0.99 (0.87)	1.00 (0.95)
Items	6	0.0 (0.74)	-0.89 to 1.38	1.00	15.22	1.01 (0.21)	1.00 (0.20)

The word 'ability' is used in the place of 'maths self-esteem' measure in order to be consistent with analyses of the other tests.

The range of student abilities was from -3.62 to 3.89 (excluding the maximum and minimum scorers), with a mean of 0.29 (SD = 1.39). The reliability of student estimates was 0.83. 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 ($\alpha = 0.86$). The student separation index was 2.25. This indicates the spread of person measures in standard error units, in this case in about 2.3 standard errors. The higher the value of this index, the more spread out the persons are on the variable being measured.

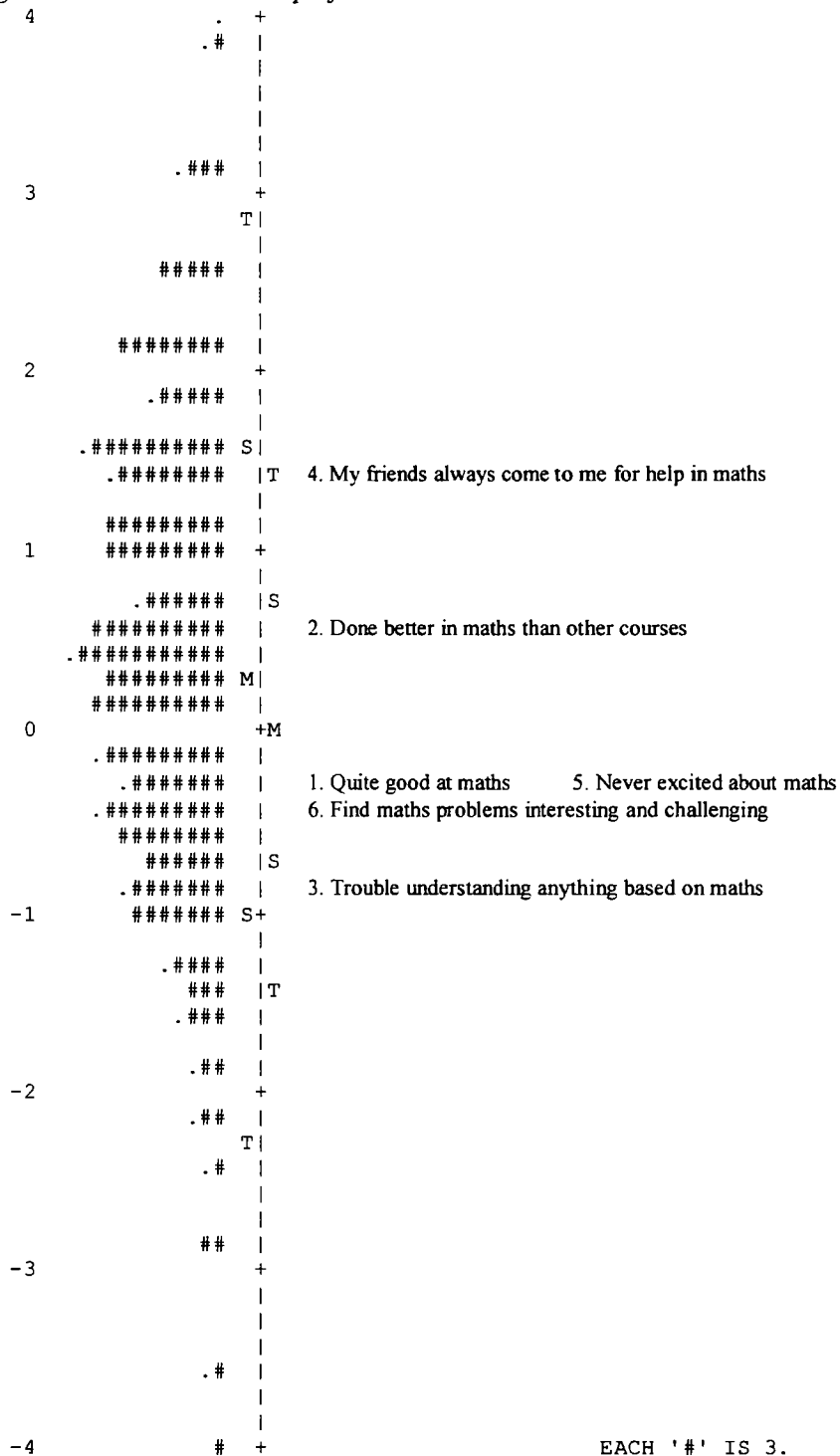
A student separation index of 2.25 also indicates approximately 3.5 statistically distinct strata ($\text{strata} = 3.33$) of student abilities identified by the instrument,

The item estimates ranged from - 0.89 to 1.38 and the reliability index was 1.00. 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 15.22, indicating that the spread of item estimates is about 15 standard errors.

There were 20 badly misfitting students ($\text{outfit and/or infit} > 3.0$), however they were not removed since the items already fitted the Rasch model well.

Figure 4.2.6 shows the item-student map. Despite the small number of items, they seem to be rather well targeted for the distribution of abilities of the students. The item measures lie between 1 standard deviation below and 1 standard deviation above the overall mean student ability. Two of the items are targeted for abilities above the overall mean ability and 4 items below.

Figure 4.2.6 Students map of items



Threshold calibrations

Table 4.2.12 below shows the bases for the estimation of the item difficulties and thresholds for the MSES.

There were 6 different categories in each item of the scale, numbered 1 – 6. The numbers corresponded to 1 = Definitely False, 2 = False, 3 = More False than true, 4 = More True than False, 5 = True, 6 = Definitely True.

The categories' columns show the response information (how many students out of the 548 used for the calibration) for each category in each item. This information has been used as the basis for the estimation of the rating scale thresholds, and that set of thresholds is applied identically to all the items on the MSES (assumption of the Rasch Rating Scale Model). In other words, the thresholds are estimated once for all items.

The 'Item Raw Score' column shows the total score for each item which has been used as the basis for the estimation of the item difficulties, which are shown in the last column (For example: Item 1 Raw Score = $1 \times 341 + 2 \times 48 + 3 \times 102 + 4 \times 144 + 5 \times 139 + 6 \times 84 = 2208$).

One can notice from the table that the item with the lowest score (item 4) has the highest measure of 1.38 (i.e. is the most difficult) and the item with the highest score (item 3) has the lowest measure of -0.89 (i.e. is the easiest).

Table 4.2.12 The bases for the estimation of item thresholds and item difficulty for the MSES

	Categories						Total	Item Raw Score	Item Measure
	1	2	3	4	5	6			
Item 1	341	48	102	144	139	84	548	2208	- 0.25
Item 2	64	100	150	102	79	53	548	1835	0.51
Item 3	18	32	64	112	165	157	548	2489	- 0.89
Item 4	152	136	111	98	40	11	548	1415	1.38
Item 5	47	70	63	117	138	113	548	2212	- 0.26
Item 6	36	50	66	115	166	115	548	2314	- 0.48
Total	348	436	556	688	727	533			

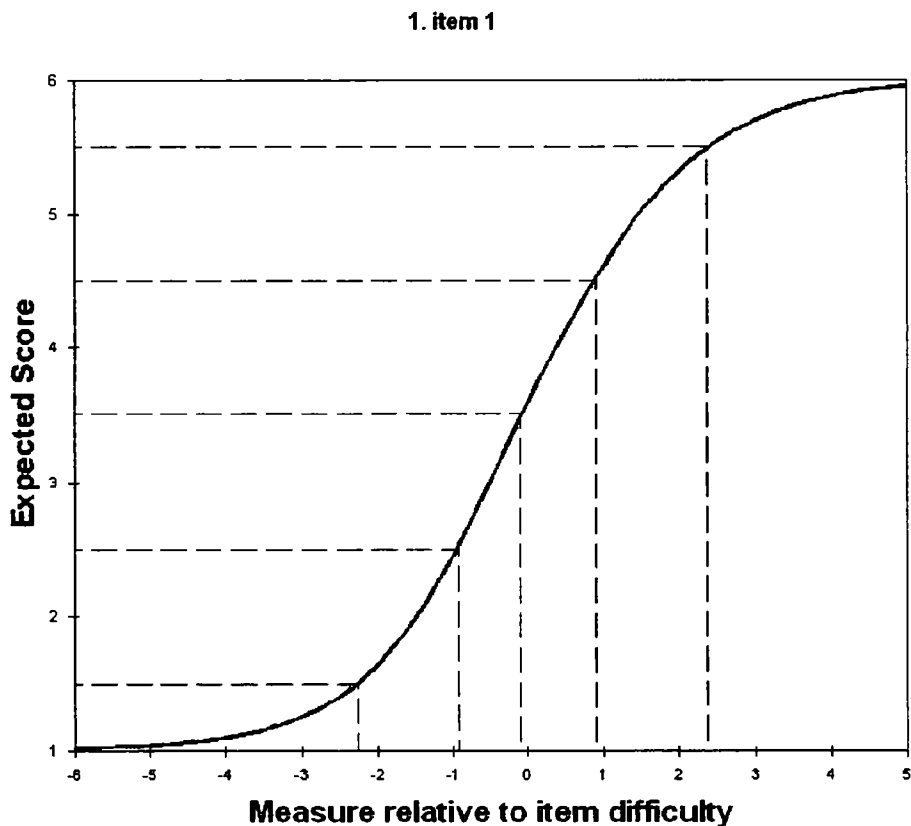
Estimating the item Thresholds

Rasch measurement provides three different approaches for estimating the thresholds, and any researcher can choose from those the one that is most meaningful for him/her.

1. Rasch-half-point thresholds.

Someone at the boundary between categories 1 and 2 would have an expected rating of 1.5. This boundary is the **expectation measure at 2 – 0.5** and is the Rasch-half-point threshold between categories 1 and 2. Similarly the threshold between categories 2 and 3 is the expectation measure at 2.5. Figure 4.2.7 below shows the expected model Item Characteristic Curve (ICC) and the expectation measures at 1.5, 2.5, 3.5, 4.5 and 5.5, and these are the Rasch-half-point thresholds.

Figure 4.2.7 Expected model ICC

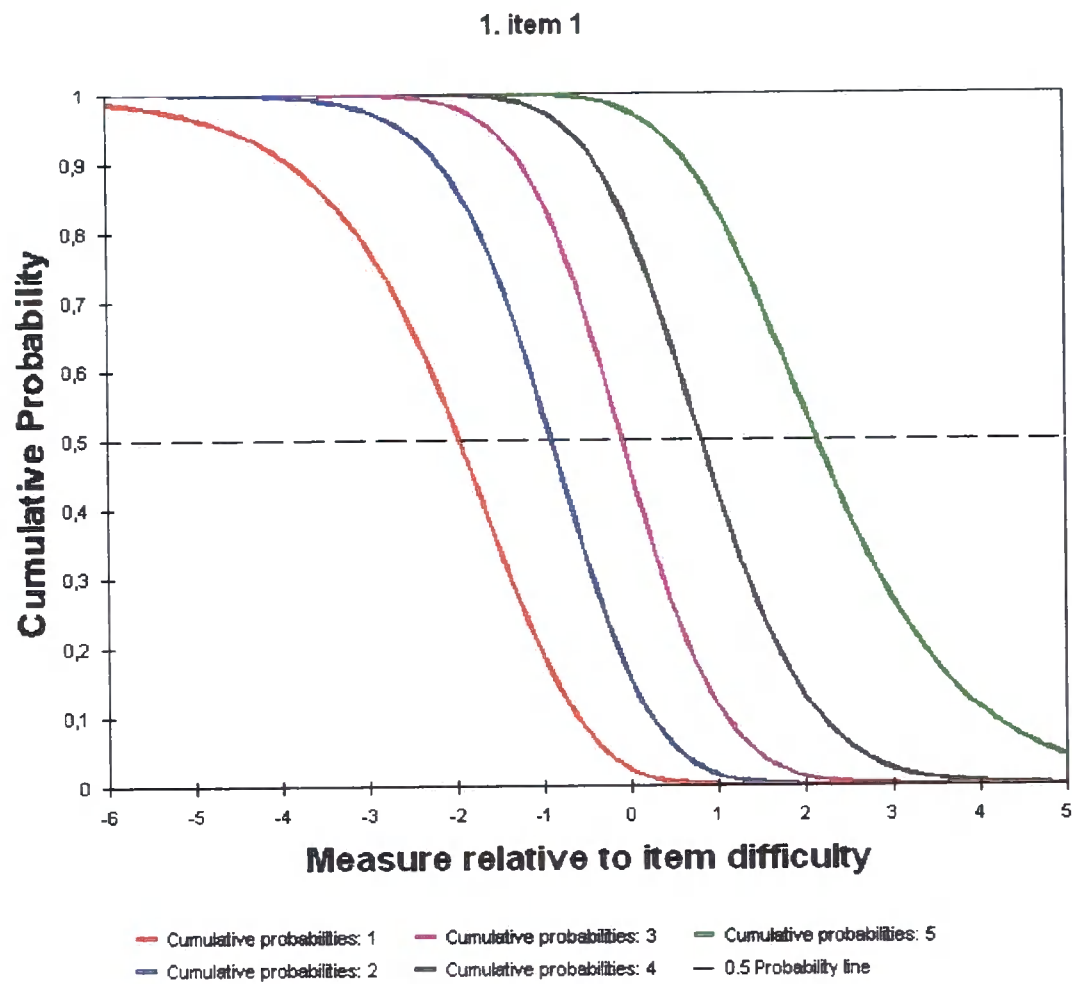


2. Rasch-Thurstone thresholds

The Rasch-Thurstone threshold between categories 1 and 2 is the measure where someone would have a 50% chance of being rated 1 or below and 50% chance of being rated 2 or above. Similarly the Rasch-Thurstone threshold between categories 4 and 5 is the measure where someone would have a 50% of being rated 4 or below and 50% chance of being rated 5 or above.

These thresholds can be seen in figure 4.2.8 which shows the cumulative probability curves. The thresholds are the measures that correspond to the points where the 0.5 line meets the cumulative probability curves for each category.

Figure 4.2.8 Cumulative probability curves



3. Rasch-Andrich thresholds

The Rasch-Andrich threshold between categories 1 and 2 is the measure at which someone has an equal chance of being rated 1 or 2. This is also called the Rasch-step calibration and it can be illustrated with the category probability curves in figure 4.2.9. The thresholds are the points of intersections of adjacent category curves and indicate when the probability of being observed in a higher category starts to exceed that of being observed into the adjacent lower one. According to Linacre (2005) this considers the categories two at a time but can lead to misinference if there is Rasch-Andrich disordering.

Figure 4.2.9 Category probability curves

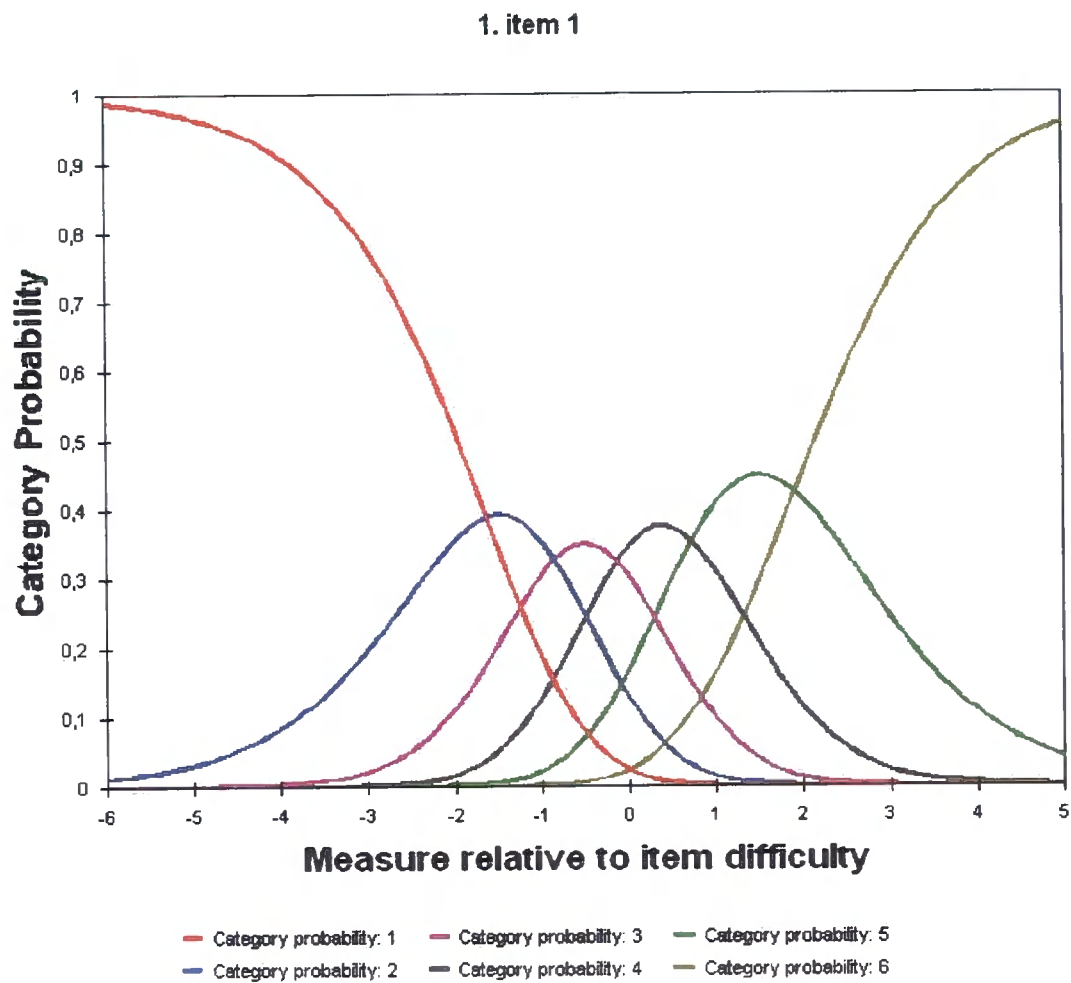


Table 4.2.13 below shows the thresholds using the 3 different approaches.

Table 4.2.13 Threshold estimations

	Categories					
	1	2	3	4	5	6
Rasch-half-point thresholds		- 2.27	- 0.95	- 0.07	0.88	2.40
Rasch-Thurstone thresholds		- 1.96	- 0.91	- 0.09	0.81	2.14
Rasch-Andrich thresholds		- 1.63	- 0.85	- 0.14	0.72	1.91

The item difficulty estimates vary from item to item, but the threshold structure modeled by the Rasch analyses of the empirical data is common to all items. The item difficulties are set as the balance points for each item and that is why in figure 4.2.10 all the item category ranges, although they are the same in widths, they differ in location according to the item difficulty.

Table 4.2.14 shows how the categories (using the Rasch-half-point thresholds) of each item are placed in the item-student map in figure 4.2.10

Table 4.2.14 Position of item categories on the student-item map

		Categories					
Items	Difficulty	1	2	3	4	5	6
1	- 0.25		-2.52	-1.2	-0.32	0.63	2.15
2	0.51		-1.76	-0.44	0.44	1.39	2.91
3	- 0.89		-3.16	-1.84	-0.96	-0.01	1.51
4	1.38		-0.89	0.43	1.31	2.26	3.78
5	- 0.26		-2.53	-1.21	-0.33	0.62	2.14
6	- 0.48		-2.75	-1.43	-0.55	0.4	1.92

Example: If the threshold between category 1 and 2 for item 4 is labeled 4.2 then:

$$\text{Measure of 4.2} = 1.38 + (-2.27) = -0.89$$

$$\text{Measure of 4.3} = 1.38 + (-0.95) = 0.43$$

$$\text{Measure of 4.4} = 1.38 + (-0.07) = 1.31$$

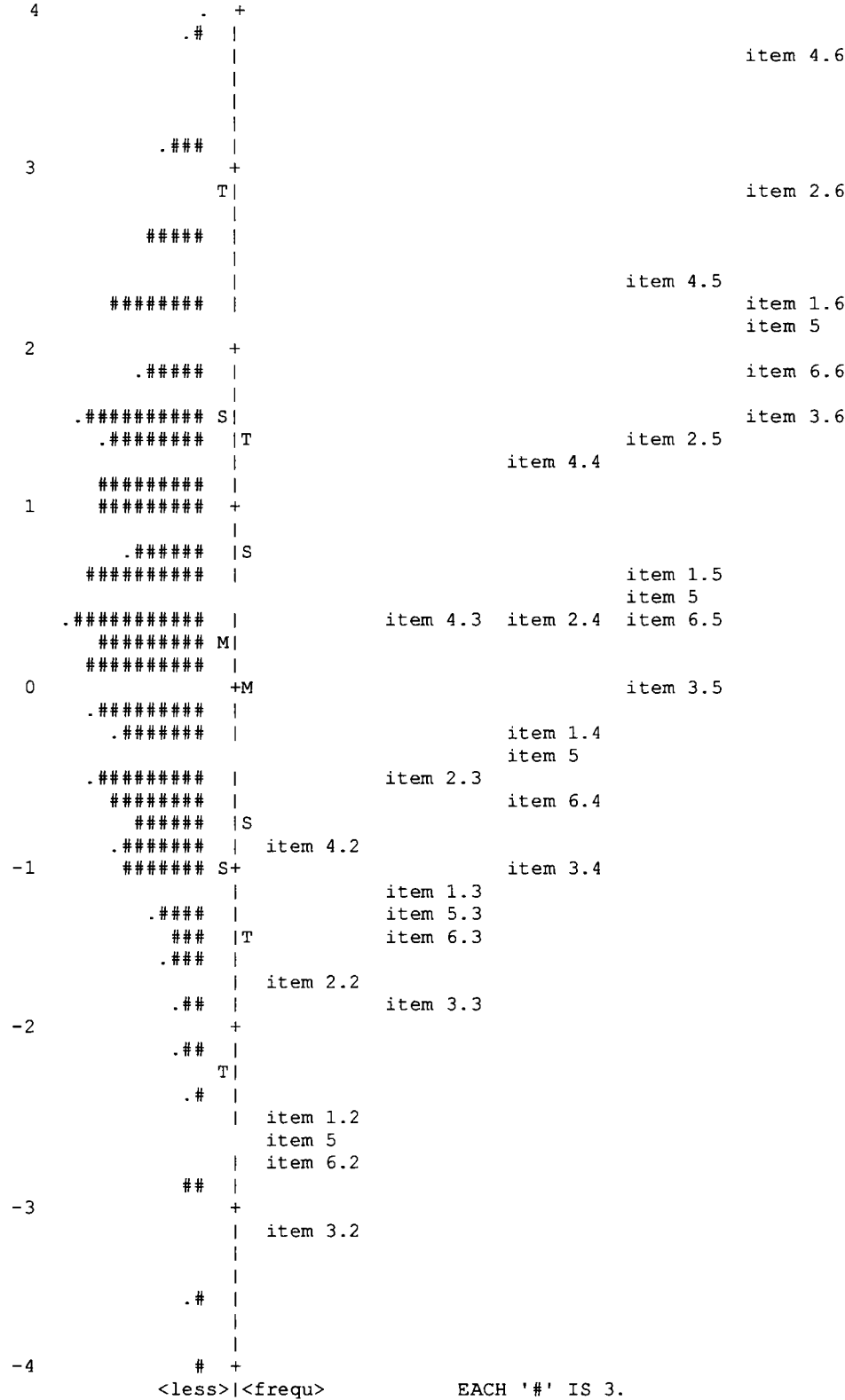
$$\text{Measure of 4.5} = 1.38 + 0.88 = 2.26$$

$$\text{Measure of 4.6} = 1.38 + 2.40 = 3.78$$

The above methods are used the same way in the case of the Partial Credit Model with the exception that since there is no common scale, i.e. each item can carry any number of marks, thresholds are calculated for each item separately.

Figure 4.2.10 shows the item – student map, this time with the thresholds. It is obvious that the various steps of the questions are well targeted for a wider range of abilities, from more than 2 standard deviations below to just more than 2 standard deviations above the overall mean ability.

Figure 4.2.10 students map of items (with item score thresholds)



As a conclusion, from the picture seen in the two figures above, one can say that despite the small number of items, these are well targeted for the distribution of student abilities and cover a wide range of abilities.

4.2.5 Reliability and validity of the MSES

For the study of the reliability of the scale two equivalent indices were used: the student reliability (this index is given by the Rasch analyses) and Cronbach's alpha. Furthermore, the item total correlations were calculated and used as another indication of the degree of internal consistency of the test.

For the validation study of the test, the following evidence was collected and presented below:

- *Principal components analysis of the raw scores*
- *Principal components analysis of the standardized residuals after the Rasch calibrations, as proposed by Linacre (1998a).*
- *A plot of the factor loadings (on the first dimension extracted, other than the dimension measured by the test) against item measures.*
- *Comparisons of the MSES scores with measures of academic achievement. These measures included the diagnostic test, the second maths test, the maths final exam and the language final exam.*
- *Male-female comparisons*

Reliability

Cronbach's alpha was used as a measure of the internal consistency of the scale. Its value was 0.862 which is very satisfactory, given the small number of items in the scale. Furthermore, all the inter-item correlations were significant (0.405 to 0.667).

Also, the reliability index, which is equivalent to the alpha but based on the measures rather than the raw scores, was calculated and it was satisfactory too (0.83).

Validity

Factor Analysis

Principal components analysis was performed using SPSS extracting only one factor. Table 4.2.15 shows the total variance explained by this factors.

Table 4.2.15 Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,582	59,699	59,699	3,582	59,699	59,699
2	,645	10,742	70,441			
3	,576	9,608	80,049			
4	,462	7,704	87,754			
5	,448	7,460	95,214			
6	,287	4,786	100,000			

Extraction Method: Principal Component Analysis.

The table suggests that the scale measures only one ability, which accounts for about 60% of the variation in the data.. The loadings of all the items on this factor are significant (from 0.686 to 0.853), supporting further the hypothesis of a unidimensional scale.

Principal components analysis of the standardised residuals

Principal components analysis on the standardised residuals (Linacre, 1988) was performed in WINSTEPS yielding:

PRINCIPAL COMPONENTS (STANDARDIZED RESIDUAL) FACTOR PLOT

Factor 1 extracts 1.5 units out of 6 units of item residual variance noise.
 Yardstick (variance explained by measures)-to-This Factor ratio: **18.1:1**
 Yardstick-to-Total Noise ratio (total variance of residuals): **4.6:1**

Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)

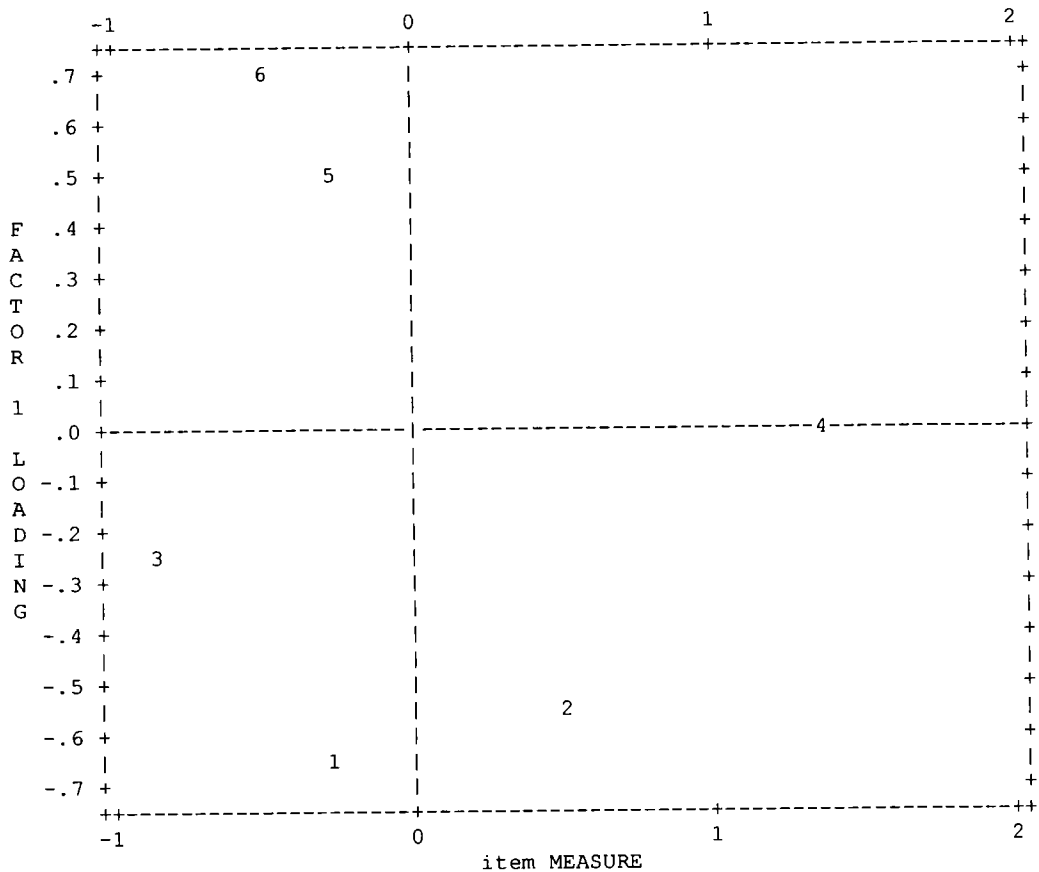
		Empirical	Modeled
Total variance in observations	=	33.6 100.0%	100.0%
Variance explained by measures	=	27.6 82.1%	81.9%
Unexplained variance (total)	=	6.0 17.9%	18.1%
Unexpl var explained by 1st factor	=	1.5 4.5%	

The variance explained by the measures (i.e. by the dimension measured by the scale) is 82.1% of the total variance. It is also more than 18 times the variance explained by the first factor extracted by PCA on the standardised residuals and about 5 times the total unexplained variance in the data. The unexplained variance is about 18% of the total variance in the data.

Also, although the variance explained by this first factor is 25% (1.5 out of 6) of the unexplained variance, which may seem high at first sight, this constitutes just 4.5% of the total variance in the data.

Given the general low stakes status of this scale (measuring a general self-esteem on maths) and the plot of loadings on the first factor extracted against the item measures given in figure 4.2.11 (where there is no indication of item groupings) one can safely conclude that there is no second dimension present in the data, therefore the scale is unidimensional.

Figure 4.2.11 Factor loadings against item measures.



Comparing correlation coefficients

Given a correlation coefficient r , it can be transformed to r^* using Fisher's transformation:

$$r^* = \text{Fi}(r) \quad \text{where} \quad \text{Fi}(r) = 0.5 \cdot \ln\left(\frac{1+|r|}{1-|r|}\right) \quad \text{Also } \text{Fi}(-r) = -\text{Fi}(r)$$

Statement

Let r be the correlation coefficient of a bivariate random sample of size n , taken from a population having correlation coefficient ρ .

Then if $r^* = \text{Fi}(r)$ and $\rho^* = \text{Fi}(\rho)$: $r^* \sim N(\rho^*, \frac{1}{n-3})$ approximately, for large n (say $n \geq 50$).

Confidence Interval for the population correlation coefficient (ρ)

From the above, one can estimate a 95% confidence interval for the transformed correlation coefficient ρ^* , from r^* using:

$$\text{Lower limit } (\rho^*_L) = r^* - 1.96 \cdot \sqrt{\frac{1}{n-3}}, \quad \text{Upper Limit } (\rho^*_U) = r^* + 1.96 \cdot \sqrt{\frac{1}{n-3}}$$

From those limits, and using the inverse Fisher's transformation $r = \frac{e^{2r^*} - 1}{e^{2r^*} + 1}$ one can estimate a 95% confidence interval for the population correlation coefficient ρ .

Comparisons with academic achievement

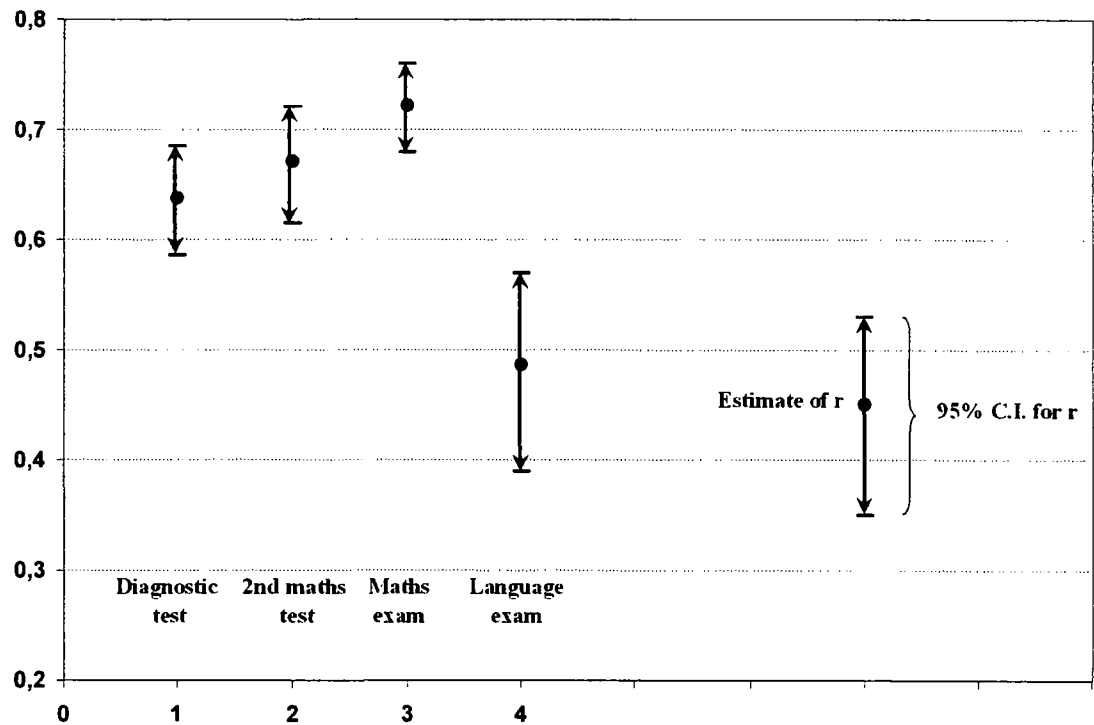
Table 4.2.16 and figure 4.2.12 below show the correlation coefficient (and its 95% confidence interval) of the scores on the MSES with

- (i) the diagnostic maths test (administered at the beginning of the year),
- (ii) the scores on the second test on quadratic equations (administered in the second term of the academic year),
- (iii) the scores on the maths final exam and
- (iv) the scores on the language final exam.

Table 4.2.16 Correlations (and their 95% confidence intervals) of the MSES scores with other criteria

	Other criteria			
	Maths diagnostic test	Second maths test	Maths final exam	Language final exam
Math self-esteem scores	0.638	0.671	0.722	486
N	553	417	540	270
r lower (95% C.I.)	0.586	0.615	0.679	0.389
r upper (95% C.I.)	0.685	0.721	0.721	0.570

Figure 4.2.12 Correlations (and their 95% confidence intervals) of the MSES scores with other criteria



It is obvious from both the table and the figure that the maths self-esteem scores correlate significantly higher with the 3 maths tests (which are considered measures of mathematical ability) than with the language exam (which is considered a measure of the language competency). These findings are consistent with Marsh's (1986) findings

that math and verbal self-concept correlate higher with the matching areas of achievement.

Male and female comparisons

Table 4.2.17 below shows the mean scores of male and female students on the MSES, the diagnostic maths test, the second maths test and the final maths exam, together with the corresponding standard deviations in brackets and the p-values of t-tests carried out for possible differences between the means.

Table 4.2.17 Scores on the MSES and maths tests or exams, by gender

Scores	Gender		p-values
	Male	Female	
MSES (max score 36)	22.94 (6.76)	22.55 (6.66)	0.492
Diagnostic maths test (max score 50)	31.04 (13.56)	33.15 (12.55)	0.046
Second test (max score 28)	12.51 (6.71)	13.32 (6.16)	0.186
Final exam (max score 20)	9.84 (6.21)	10.73 (5.97)	0.074

Apart from the diagnostic test, where female students scored significantly higher than male students, there are no significant differences in the other tests or the MSES. However, it is worth noting that in both the maths tests and the maths exam the female students scored higher, but the male students seem to have slightly higher self-esteem in maths.

Similarly all standard deviations were higher for the males' scores with only the standard deviations in the diagnostic maths test differing significantly ($p = 0.019$, using the F-test).

These findings are consistent with studies by many researchers (see Marsh et. al (1988); Skaalvik and Skaalvik, (2004)) who found that male students had higher self-concept in maths, meaning that males seem to judge themselves more favourably than females do. However, none of the gender differences in maths self-concept could be explained by differences in achievement.

All of the above evidence collected leads to the conclusion that the MSES has a high degree of validity.

Different self-esteem groups

The range of abilities (self-esteem measures) was divided into three different groups, the low, medium and top anxiety groups using three different cut-off scores, the 30th and 70th percentiles, the 20th and 80th percentiles and the 10th and 90th percentiles.

Misfitting students

Following the calibration of the MSES, misfitting students were identified using cut-off scores for the infit and outfit mean squares of 1.5. Table 4.2.18 shows the number of students identified as misfitting by the two indices as well as the total number.

*Table 4.2.18 Misfit (infit) * Misfit (outfit) Crosstabulation*

		Misfit (outfit)		Total
		Fitting	Misfitting	
Misfit (infit)	Fitting	453	9	462
	Misfitting	13	78	91
Total		466	87	553

The number of students identified as misfitting by the infit statistic was 91 (16.5%) and by the outfit statistic was 87 (15.7%) whereas 78 students were identified by both, giving a total of 100 (18.1%) misfitting students.

4.2.6 The second maths test (test on quadratic equations)

The sample

The test was administered to 445 out of the original 635 students taking the diagnostic test. In Limassol 1, 11 classes and 266 students were involved (one teacher teaching one class did not want to participate and some students were absent on the day of the test). In Limassol 2, 2 classes and 37 students (a few students were absent), and in Paphos, 6 (out of the original 11) classes and 142 students (2 teachers teaching 5 classes did not want to participate and a few students were absent). A total of 19 classes were involved. Overall, out of the total of 445 students, 41.8% were male (similar to the 43.9% in the first test) and 58.2% female (similar to the 56.1% in the first test).

Table 4.2.19 shows the distribution of the 445 students by gender, in the three different schools. In all the schools the proportion of female students was again larger than that of male students.

Table 4.2.19 School * Gender Crosstabulation

		School			Total
		Limassol 1	Limassol 2	Paphos	
Gender	Male	104	15	67	186
	Female	162	22	75	259
Total		266	37	142	445

Test calibrations

The Rasch model was used for the calibrations. The first calibration on the full dataset revealed only one misfitting item (outfit = 1.74) and 23 badly misfitting students (outfit > 3.0).

The 23 students were removed and a second calibration was performed, revealing only 2 marginally misfitting items (outfit values of 1.37 and 1.33). Those items were retained in the dataset (the reasons for not removing the items are explained).

The item statistics from the second calibration were then used for the final calibration in order to get the students statistics.

Item-person maps are presented to show how well the items were targeted for the population of students and finally the students were divided into groups according to their ability for investigating later on whether ability is associated with misfit.

First calibration

The first calibration, in which the full set of the test data was used (16 items of which 12 were multiple choice items giving 1 mark for the correct answer and 0 marks for an incorrect answer and 445 students), revealed one misfitting item, item 13 (outfit = 1.74) and a couple of marginal items, items 10 and 2 (outfit = 1.28) as shown in table 4.2.20. The mean values of infit and outfit were 0.99 and 1.08 respectively.

Table 4.2.20 Item statistics: misfit order

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	ERROR	INFIT		OUTFIT		PTMEA CORR.	items
					MNSQ	ZSTD	MNSQ	ZSTD		
13	1378	437	-1.07	.06	.98	-.2	1.74	3.7	A .65	item 13
10	177	437	.72	.11	1.10	2.0	1.28	3.2	B .41	item 10
2	244	437	-.09	.11	1.17	3.5	1.28	3.5	C .37	item 2
1	375	437	-2.09	.15	.94	-.6	1.21	1.0	D .39	item 1
8	183	437	.65	.11	.98	-.5	1.20	2.4	E .49	item 8
6	280	437	-.54	.11	1.08	1.6	1.10	1.1	F .42	item 6
9	167	437	.85	.11	1.05	1.1	1.09	1.0	G .45	item 9
11	187	437	.60	.11	1.02	.6	1.06	.8	H .47	item 11
15	337	437	1.51	.06	1.04	.5	1.01	.1	h .67	item 15
16	303	437	1.68	.06	1.03	.3	.99	.0	g .66	item 16
7	230	437	.08	.11	.98	-.4	.98	-.3	f .50	item 7
5	374	437	-2.07	.15	.92	-.8	.94	-.2	e .43	item 5
3	339	437	-1.39	.13	.94	-.8	.79	-1.5	d .48	item 3
4	187	437	.60	.11	.90	-2.3	.88	-1.5	c .55	item 4
12	268	437	-.39	.11	.89	-2.2	.86	-1.7	b .55	item 12
14	551	437	.95	.05	.86	-1.8	.80	-2.0	a .75	item 14
MEAN	349.	437.	.00	.10	.99	.0	1.08	.6		
S. D.	283.	0.	1.13	.03	.08	1.5	.23	1.8		

Table 4.2.21 shows the top part of the table with the student statistics from the original calibration in misfit order (outfit > 2.30 and/or infit > 2.30).

The first 23 students in table 4.2.18 with outfit (or infit) > 3.0 (5.2%) were considered badly misfitting and distorting the calibration process.

Table 4.2.21 Student Statistics: MISFIT ORDER

ENTRY	RAW					INFIT	OUTFIT	PTMEA	
NUMBER	SCORE	COUNT	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	ZSTD	CORR.
261	12	16	.22	.48	6.28	4.3	4.92	4.8	A-.02
429	10	16	-.26	.49	4.47	3.6	6.05	5.5	B .12
304	12	16	.22	.48	5.26	3.8	3.74	3.8	C .52
340	12	16	.22	.48	5.16	3.8	3.66	3.7	D .51
30	27	16	3.21	.95	.98	.5	4.97	1.8	E-.24
355	17	16	1.12	.37	2.74	2.6	4.50	3.6	F .35
634	24	16	2.09	.44	.67	-.3	4.48	2.3	G-.09
544	17	16	1.12	.37	2.74	2.6	4.27	3.4	H .12
81	23	16	1.92	.40	1.22	.6	4.14	2.3	I-.02
41	13	16	.44	.46	4.05	3.0	2.65	2.6	J .34
179	12	16	.22	.48	3.88	3.0	2.88	2.9	K .48
368	20	16	1.51	.35	1.63	1.5	3.80	2.6	L .29
221	26	16	2.62	.64	.96	.3	3.61	1.7	M-.08
114	10	16	-.26	.49	3.54	2.9	2.53	2.5	N .46
129	21	16	1.63	.36	.87	-.2	3.30	2.2	O-.01
282	18	16	1.25	.36	2.45	2.5	3.30	2.6	P .23
363	1	16	-3.30	1.06	1.24	.6	3.30	1.4	Q-.15
586	1	16	-3.30	1.06	1.24	.6	3.30	1.4	R-.15
477	22	16	1.77	.37	.39	-1.6	3.24	2.0	S .04
37	8	16	-.72	.47	1.37	.8	3.18	2.9	T .56
115	8	16	-.72	.47	3.15	2.9	2.35	2.1	U .02
296	26	16	2.62	.64	.91	.3	3.14	1.5	V .07
143	20	16	1.51	.35	.58	-1.1	3.07	2.2	W .02
65	20	16	1.51	.35	1.22	.7	2.80	2.0	X .11
96	18	16	1.25	.36	.82	-.3	2.75	2.2	Y .06
295	17	16	1.12	.37	2.67	2.5	2.74	2.3	Z .49
555	13	16	.44	.46	2.73	2.1	2.00	1.8	.53
176	16	16	.97	.39	1.34	.7	2.56	2.2	.14
177	12	16	.22	.48	2.53	2.0	2.28	2.2	.44
328	16	16	.97	.39	2.08	1.7	2.36	2.0	.29
369	8	16	-.72	.47	2.30	2.0	1.23	.6	.57
161	10	16	-.26	.49	1.64	1.1	2.30	2.2	.35

These 23 students were therefore removed, leading to a second calibration with again the 16 items, but this time with 422 students.

Second calibration

Table 4.2.22 below shows the item statistics from this second calibration. This time, the marginal items remain marginal (item 10: outfit = 1.37 and items 2: outfit = 1.33). However item 13 is now fitting the model very nicely with an outfit value of 0.99 and an infit value of 0.80.

The mean value of the infit and the outfit are 0.99 (as before) and 1.00 respectively.

Table 4.2.22 Item statistics: misfit order

ENTRY	RAW				INFIT		OUTFIT		PTMEA	
NUMBER	SCORE	COUNT	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	ZSTD	CORR.	items
10	165	414	.75	.12	1.13	2.5	1.37	3.8	A .40	item 10
2	232	414	-.12	.11	1.20	3.7	1.33	3.8	B .37	item 2
6	265	414	-.57	.12	1.12	2.2	1.17	1.8	C .41	item 6
9	157	414	.86	.12	1.09	1.7	1.15	1.6	D .44	item 9
11	176	414	.61	.11	1.07	1.5	1.13	1.5	E .45	item 11
16	265	414	1.82	.06	1.05	.5	.95	-.2	F .66	item 16
7	217	414	.07	.11	1.02	.4	1.04	.6	G .49	item 7
1	359	414	-2.23	.16	.92	-.7	1.02	.1	H .40	item 1
8	170	414	.69	.11	.97	-.6	1.00	.1	h .51	item 8
13	1328	414	-1.20	.06	.80	-2.2	.99	.0	g .68	item 13
15	283	414	1.69	.06	.97	-.3	.70	-1.6	f .68	item 15
3	321	414	-1.43	.13	.96	-.5	.83	-1.1	e .47	item 3
4	175	414	.62	.11	.91	-1.9	.90	-1.2	d .55	item 4
12	253	414	-.40	.12	.91	-1.8	.89	-1.3	c .55	item 12
5	358	414	-2.20	.16	.90	-1.0	.72	-1.3	b .44	item 5
14	496	414	1.05	.05	.87	-1.7	.76	-2.5	a .76	item 14
MEAN	326.	414.	.00	.11	.99	.1	1.00	.2		
S.D.	273.	0.	1.21	.03	.11	1.7	.19	1.8		

Both of the slightly misfitting items were multiple-choice, dichotomously scored items. Further investigation was conducted into the marginal misfit of these items.

Table 4.2.23 below shows the number (and percentage) of students scoring 0 or 1 mark in these two questions. Table 4.2.24 shows the most unexpected answers to those two questions and the students giving those answers, in ability order (highest to lowest, from left to right). The students' entry numbers are shown in columns.

Table 4.2.23 Score frequencies and percentages for items 10 and 2.

Items	Score	Number of students	Percentage
10 (Measure = 0.75)	0	250	59
	1	172	41
2 (Measure = - 0.12)	0	183	43
	1	239	57

Table 4.2.24 Most misfitting response strings

item	OUTMNSQ	student
		1 52 223 642 232164221 55511531212 55433321552 5
		1095835806264992972876789657645697582 7617609488545
		182663373553465886331657165760613048730400669483969
	high	-----
10 item 10	1.37 A	0..0.....111...1.11.1.11..
2 item 2	1.33 B	..0.....0..0.....1...1..1..1..

Item 10 was one of the rather harder items with a measure of 0.75. It was the second hardest from the 12 multiple-choice items (item 9 was harder with a measure of 0.86). The question was:

“Given that the discriminant of the quadratic equation $ax^2 + bx + c = 0$ is 20, what is the discriminant of the quadratic equation $cx^2 + bx + a = 0$?”

The marginal outfit value of this item (1.37) was mainly caused by a few unexpectedly correct answers by low ability students, probably by guessing. The most unexpected of these answers were given by students with entry numbers 254, 88, 27, 410, 366, 306, 144, 583 and 259.

Students with entry numbers 27, 306 and 259 were very low ability students. Their ability estimates were - 1.76, - 2.09 and - 2.56 respectively.

Item 2 was one with about average difficulty (measure of - 0.12). The question was:

“Given the quadratic equation $3x - 2x^2 - 5 = 0$, state the values of a, b and c” (a, is the coefficient of x^2 , b the coefficient of x and c the constant of the trinomial).

Because the quadratic equation was not given in the usual order ($ax^2 + bx + c = 0$, in descending powers of x), a few students got confused and gave a wrong answer.

The marginal outfit value of this item (1.33) was caused by a few unexpectedly wrong (by higher ability students) and a few unexpectedly correct answers (by lower ability students).

For example, the students with entry number 556 (ability estimate of 2.76) and 625 (ability estimate of 1.35) gave an incorrect answer, whereas students with entry numbers 3 (ability estimate of – 1.76) and 259 (ability estimate of – 2.56) found the correct answer.

Since the two items were marginally misfitting, and that was caused by few unexpected responses, they were not removed from the calibration process.

A summary of the results of the Rasch analysis from the 2nd calibration is given in table 4.2.25

Table 4.2.25 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)
Examinees	422	0.25 (1.29)	-3.36 to 3.34	0.83	2.23	0.96 (0.49)	1.00 (0.56)
Items	16	0.0 (1.21)	-2.23 to 1.82	0.99	10.53	0.99 (0.11)	1.00 (0.19)

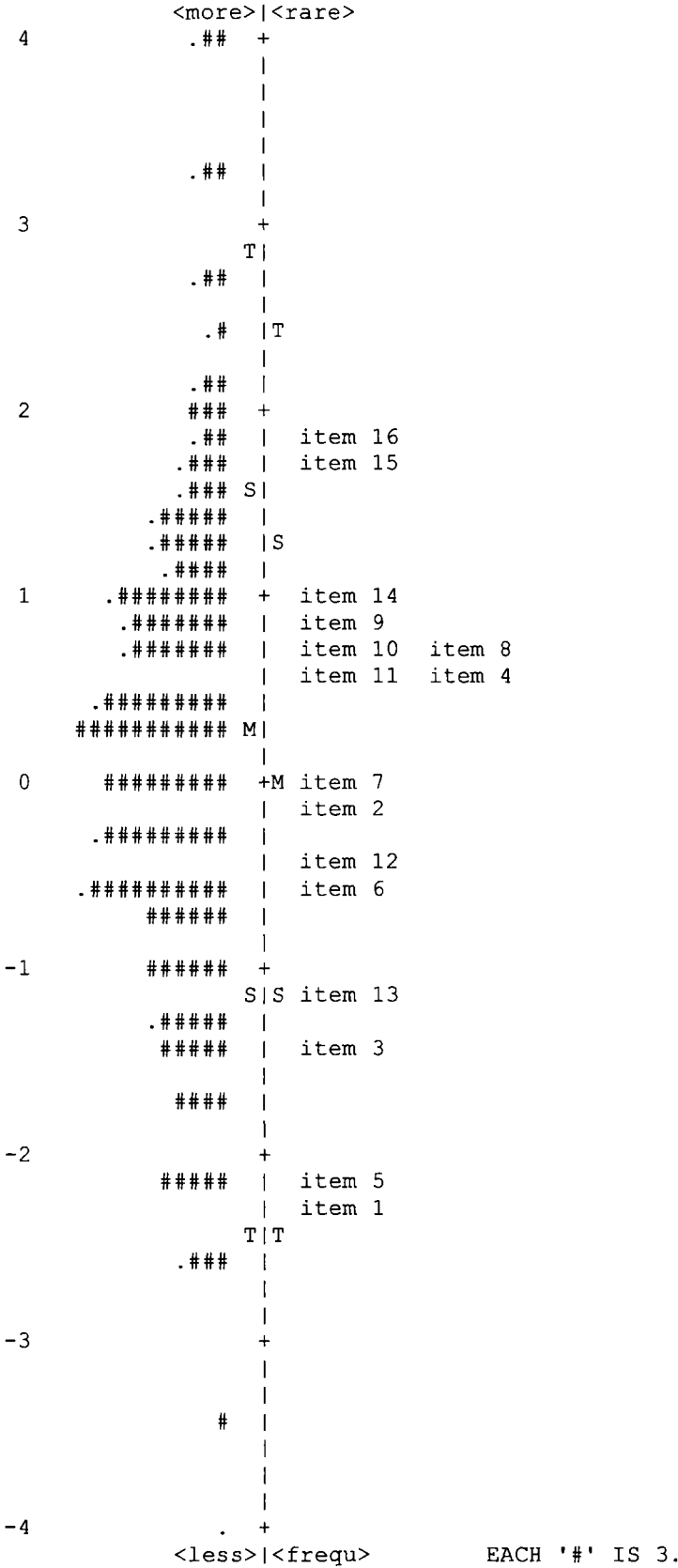
The range of student abilities was from -3.36 to 3.34, with a mean of 0.25 (SD = 1.41). The reliability of student estimates was 0.83 and it is equivalent to Cronbach’s alpha (alpha = 0.81). The student separation index was 2.23 and it indicates that the spread of person measures was about 2.2 standard errors. The higher the value of this index, the more spread out the persons are on the variable being measured. A student separation index of 2.23 also indicates approximately 3.5 statistically distinct strata (strata = 3.31) of student abilities identified by the instrument.

The item estimates ranged from - 2.23 to 1.82 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 10.53, indicating that the spread of item estimates is about 11 standard errors.

Third and final calibration

The statistics of the items from the second calibration were then used for the third and final calibration which included the 16 anchored items and all the 445 students. Figure 4.2.13 shows the item-student map.

Figure 4.2.13 Students map of items



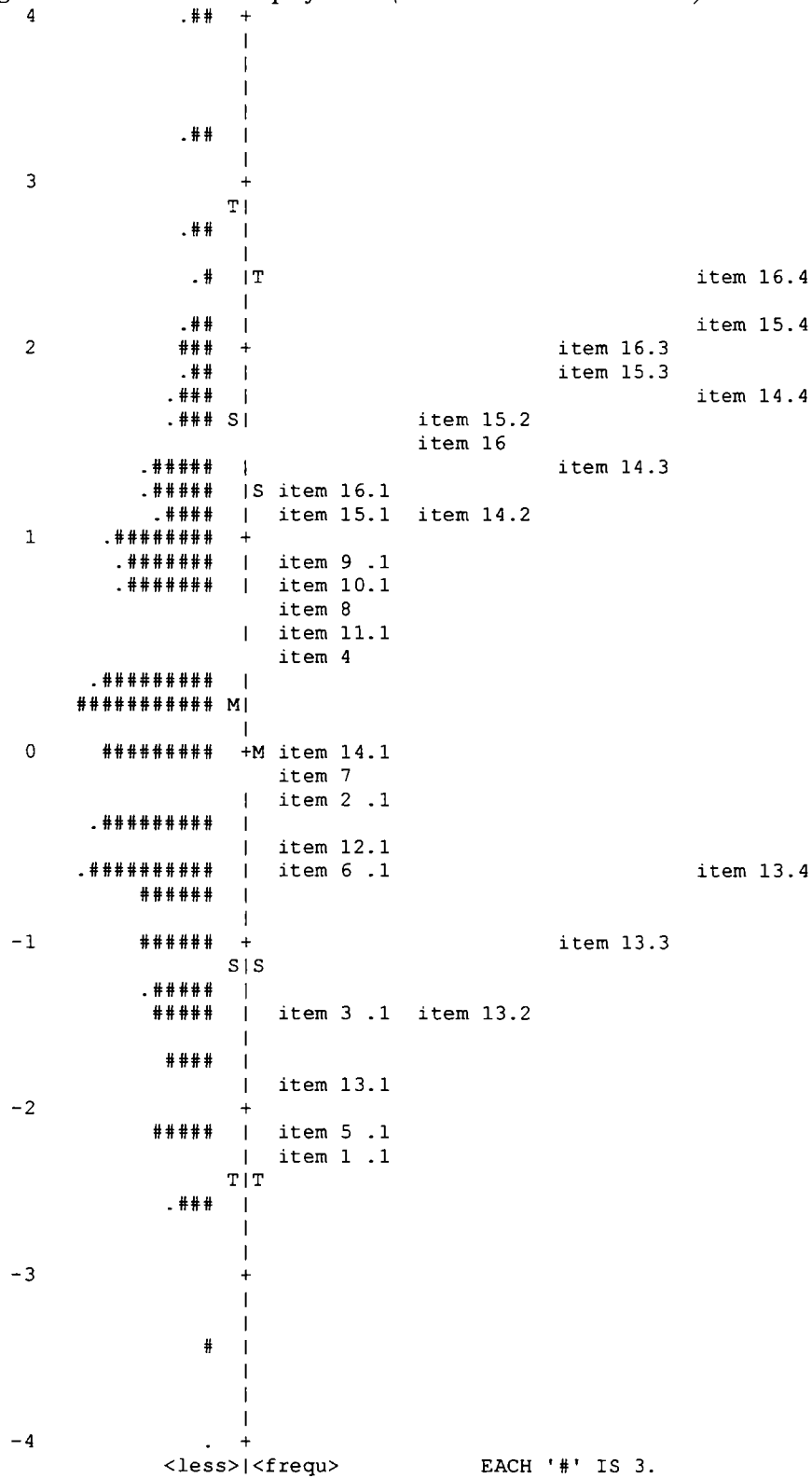
The distributions of item difficulties and students' abilities are almost symmetrical indicating a very well designed test. The items are targeted for students with abilities from 2 standard deviations below to about one and a half standard deviations above the mean student ability.

Also 8 items have difficulties above the students' mean ability and 8 items below.

Figure 4.2.14 shows another item – student map, this time with all the categories of the items (the thresholds for all the possible scores for each item). The first 12 items are dichotomously scored and only items 13 to 16 have three thresholds (they carry 4 marks each).

It is obvious that the various steps of the questions are well targeted for a wider range of abilities, from 2 standard deviations below to about two standard deviations above the overall mean ability, covering approximately the central 95% of the distribution of abilities.

Figure 4.2.14 students map of items (with item score thresholds)



4.2.7 Reliability and validity of the 2nd test

For the study of the reliability of the test two equivalent indices were used: the student reliability (this index is given by the Rasch analyses) and Cronbach's alpha.

For the validation study of the test, the following evidence was collected and presented below:

- *Analysis of a content validity questionnaire*
- *Principal components analysis of the standardized residuals after the Rasch calibrations, as proposed by Linacre (1998a).*
- *A plot of the factor loadings (on the first dimension extracted, other than the dimension measured by the test) against item measures.*
- *Correlations of the maths test scores with the final maths exam scores.*
- *Comparisons of the item estimates from two different calibrations (based again on students' gender) to ascertain whether invariance holds.*
- *Comparisons of ability estimates from the two maths tests used in this phase of the study.*

Reliability

The student reliability was 0.83 and Cronbach's alpha was (0.81).

These measures are not as high as the equivalent measures in the other two tests (one in phase 1: student reliability = 0.86, alpha = 0.91 and the diagnostic test in phase 2: student reliability = 0.87, alpha = 0.90). 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 2.23. A student separation index of 2.23 also indicates approximately 3 and a half statistically distinct strata (strata = 3.31) of student abilities identified by the instrument.

Validity of the test

The same questionnaire that was used in phase 1 of this study, on content validity, was again administered, this time to 8 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 included any items on skills not included in the syllabus.

Table 4.2.26 shows the number of experts who selected each option in each of the six statements.

Table 4.2.26 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

It is clear that all the experts agree or absolutely agree on almost all the statements regarding the content validity of the test.

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 worry as to the time limits, arguing that the questions were probably too many to be answered within a 45-minute period. However, the administration of the test proved that there was no problem with the time given to the students to complete the test.

Principal components analysis of the standardised residuals

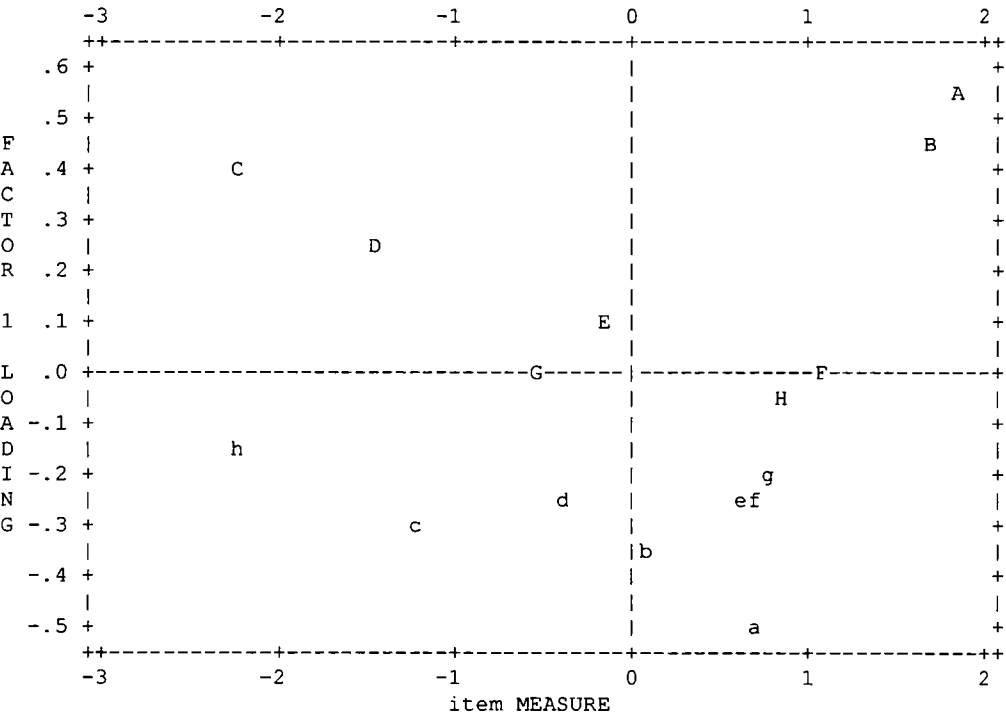
Principal components analysis on the standardised residuals (Linacre, 1988) was performed in WINSTEPS yielding:

PRINCIPAL COMPONENTS (STANDARDIZED RESIDUAL) FACTOR PLOT				
Factor 1 extracts 1.5 units out of 16 units of item residual variance noise.				
Yardstick (variance explained by measures)-to-This Factor ratio: 91.7:1				
Yardstick-to-Total Noise ratio (total variance of residuals): 8.4:1				
Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)				
		Empirical		Modeled
Total variance in observations	=	150.7	100.0%	100.0%
Variance explained by measures	=	134.7	89.4%	89.3%
Unexplained variance (total)	=	16.0	10.6%	10.7%
Unexpl var explained by 1st factor	=	1.5	1.0%	

The variance explained by the measures (i.e. by the dimension measured by the test) is 89.4% of the total variance. It is also about 92 times the variance explained by the first factor extracted by PCA on the standardised residuals and about 8.5 times the total unexplained variance in the data. The unexplained variance is 10.6% of the total variance in the data.

Also, the variance explained by this first factor is about only 9.4% (1.5 out of 16) of the unexplained variance and just 1% of the total variance in the data. Given these results and the plot of loadings on the first factor extracted against the item measures given in figure 4.2.15 where there is no indication of item groupings one can safely conclude that there is no second dimension present in the data, therefore the test is unidimensional.

Figure 4.2.15 Factor 1 loadings against item measures.



Correlations with the final maths exams

The scores on the test were compared with the final mathematics exam results of the students in the 3 schools. This was done separately for each school since each school prepared its own final examination. The correlation coefficients (all highly significant) were:

- Limassol 1: $r = 0.840$ ($N = 259$)
- Limassol 2: $r = 0.634$ ($N = 36$)
- Paphos: $r = 0.751$ ($N = 141$)

The total number of students adds up to 436 (instead of the original 445) because 9 of the students who took the test were either asked to take the exams in September, or to repeat the year, because of too many unauthorised absences.

Finally, the correlation of the scores on this test with the scores on the diagnostic test which took place at the beginning of the year was 0.711 ($p < 0.01$) showing a highly significant positive correlation, but perhaps a little low most probably because of the

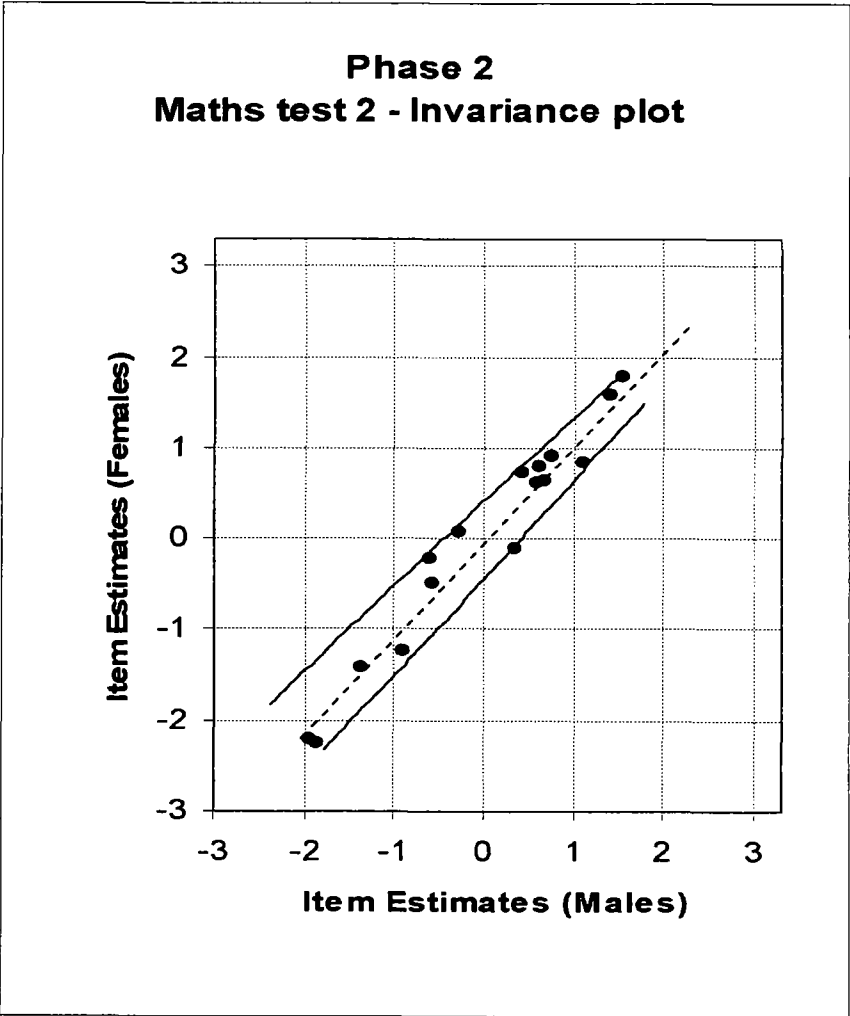
different targeting of the items in the two tests. The diagnostic was targeted for the low ability students (and spread from mean ability – 2,5 SD to mean ability + 0,5 SD) whereas test 2 for the average ability students (spread from mean ability – 2 SD to mean ability + 2 SD)

Comparisons of item estimates from two calibrations

Split of the data by gender

In this case the data was split into two groups based on gender. The two groups had sizes 186 (males) and 259 (females). Figure 4.2.17 below shows the invariance plot for the item estimates from these two subsets.

Figure 4.2.17 Invariance plot for the second maths test(by gender)



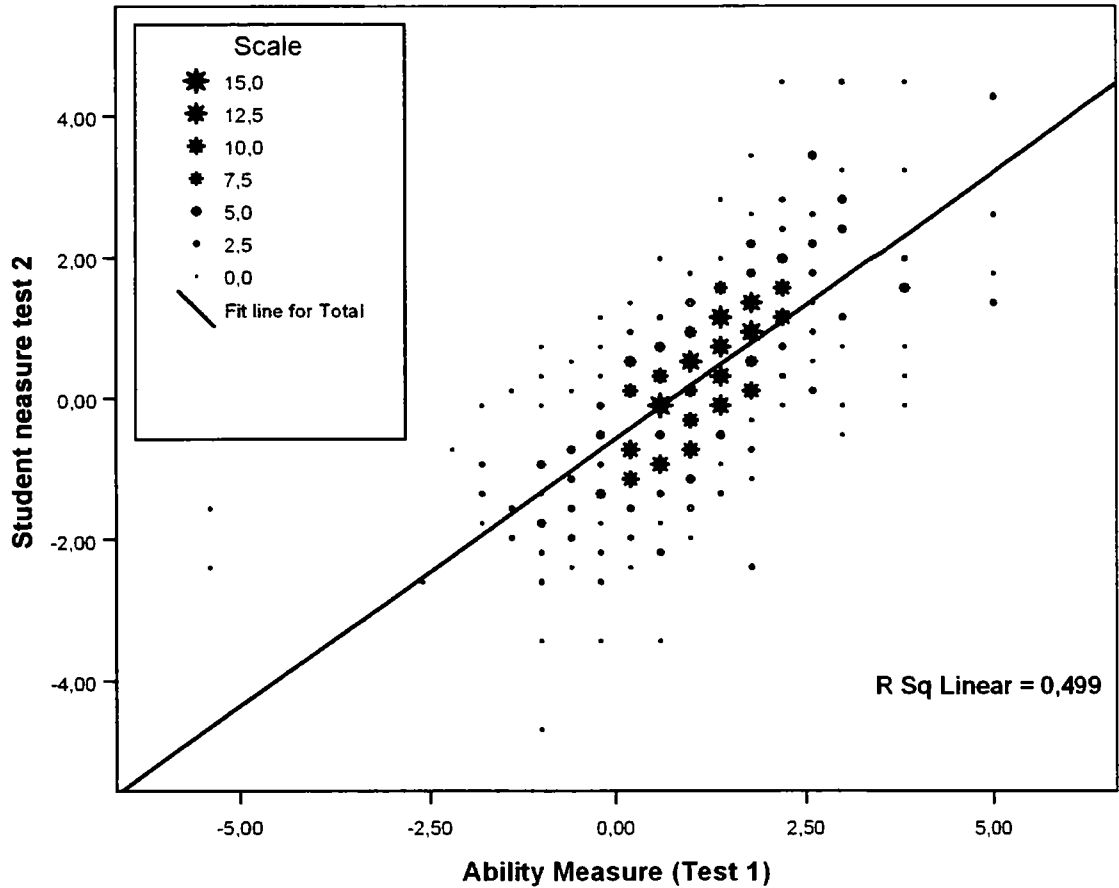
The points are closely scattered around the identity line, with no items outside the confidence limits, and that is a strong indication that invariance holds. Also, the correlation coefficient is 0.979 which is extremely high.

This invariance of item calibrations across groups supports the hypothesis that the construct measured by this instrument has the same meaning to the groups which were studied.

Comparing ability estimates from calibrations of two different tests

The students' ability estimates from this test were compared with the ability estimates from the diagnostic test. Figure 4.2.18 shows the scatter diagram with the line of best fit of the ability estimates from test 2 against the ability estimates from the diagnostic test (test 1).

Figure 4.2.18 Scatter diagram of ability estimates (test 2 against test 1)



The correlation coefficient of this comparison was 0.706, and was highly significant.

This strengthens 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, they were targeted at different ability-level students. The first test was very easy, and being a diagnostic test aiming to investigate whether the students had the basic mathematical skills required for the first form of the lyceum, was targeted for the lower ability students. The second test was targeted for 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.

Ability groups

The range of abilities was again divided into three different groups, the low, medium and top ability groups using the same three different cut-off scores.

First, the range of abilities was divided into 3 groups using the 10th (measure of -1.486) and 90th (measure of 1.889) percentiles.

Second, the range of abilities was divided into 3 groups using the 20th (measure of -1.008) and 80th (measure of 1.352) percentiles.

Finally, the range of abilities was divided into 3 groups using the 30th (measure of -0.526) and 70th (measure of 1.054) percentiles.

The groups formed were labelled Low, medium and top ability groups.

4.2.8 Misfitting students

Misfitting students were identified using appropriate cut-off scores for the infit and outfit statistics (1.3 for both). The numbers and proportions of misfitting students are presented, together with comparisons of equivalent proportions from a simulation study.

Hence, an investigation was carried out into whether the same students misfit in administrations of different maths tests.

Following the calibration of the test, misfitting students were identified using cut-off scores for the infit and outfit mean squares of 1.3.

Table 4.2.28 shows the number of students identified as misfitting by the two indices as well as the total number.

Table 4.2.28 Misfit (infit) * Misfit (outfit) Crosstabulation

		Misfit (outfit)		Total
		Fitting	Misfitting	
Misfit (infit)	Fitting	308	41	349
	Misfitting	32	64	96
Total		340	105	445

The number of students identified as misfitting by the infit statistic was 96 (21.6%) and by the outfit statistic was 105 (23.6%) whereas 64 students were identified by both, giving a total of 137 (30.8%) misfitting students.

A simulation study was carried again out using WINSTEPS (Linacre, 2005).

The infit mean square calculated for this Rasch-fitting data set identified 12.9% misfitting students (infit > 1.3) and the outfit mean square 19.1 % (outfit > 1.3). The proportion of misfitting students in the simulated dataset was 24.7%.

The results of the simulation study show a similar proportion for the outfit and a lower proportion for the infit than the results from the analyses of the test data. The overall proportion of misfitting students in the simulated data is slightly lower than that in the actual test data.

However, given the fact that the simulated data fit the Rasch model perfectly, and that 2 marginally misfitting items were retained in the test, the comparisons do suggest a reasonable fit of the data to the Rasch model.

Do the same students misfit in different maths tests?

The next table, table 4.2.29 shows the numbers of fitting and misfitting students in the diagnostic maths test and test 2, for the purpose of testing whether there is any association between them.

It shows that 31.9% (92 out of 288) of the fitting students in the diagnostic test and 28.7% (45 out of 157) of the misfitting students in the diagnostic test were also misfitting in test 2.

Similarly, 36.4% (112 out of 308) of the fitting students in test 2 and 32.8% (45 out of 137) of the misfitting students in test 2 were also misfitting in the diagnostic test.

Both of these results are not significant indicating that the proportions of misfitting students in the second test are similar for the fitting and misfitting groups of students in the diagnostic test. That is, there is no association between misfitting in the first and the second maths tests.

*Table 4.2.29 Misfit in test 1 * Misfit in test 2 Crosstabulation*

		Misfit in test 2 (On quadratic equations)		Total
		Fitting Students	Misfitting Students	
Misfit in test 1 (Diagnostic test)	Fitting Students	196	92	288
	Misfitting Students	112	45	157
Total		308	137	445

Chi-square = 0.371, d.f. = 1, p = 0.542

4.2.9 The shorter version of the TAI

The sample

The TAI was administered to 504 of the 635 students taking part in the study. All of these 504 students had taken the first maths test, the diagnostic one. However, only 383 of these had taken both the TAI and the second maths test on the quadratic equations.

Constructing the short TAI

The shorter version of TAI was developed from the analyses of the original one administered in phase 1 and consisted of 10 items, aiming to measure the overall test anxiety of the respondents.

Out of the 8 items measuring the worry factor in the original TAI, 4 (the items with the highest loadings on the worry factor) were selected. Similarly, out of the 8 items measuring the emotionality factor, 4 were selected, again the ones with the highest loadings on the emotionality factor.

Finally, from the 4 remaining items on the original scale, which measure general anxiety, 2 were selected based on their infit and outfit values. The two items with mean square statistics closer to 1, the expected value of these statistics according to the Rasch model, were selected.

Table 4.2.30 shows the statements selected from the original TAI and used for the shorter version, together with their loadings on the two factors (columns 2 and 3) and their measure (column 4), infit values (column 5) and outfit values (column 6) from the original analyses in phase 1.

The items measuring the Emotionality factor are shown by the highlighted loadings under the emotionality column and the items measuring Worry by the highlighted loadings under the worry column. The two items that measure total anxiety (items 13 and 19) are the ones with no highlighting in their loadings. They can be identified by the highlighting in their infit and outfit values, because those values were the criterion used for their selection.

Table 4.2.30 The shorter version of the TAI

Statement	Anxiety factors		Rasch analyses		
	Emot.	Worry	Meas.	Outfit	Infit
1. Thoughts of doing poorly interfere with my concentration on tests.	0.572	0.624	- 0.16	0.85	0.86
2. I feel very jittery when taking an important test.	0.728	0.446	0.23	0.84	0.88
3. During tests I feel very tense.	0.783	0.523	- 0.09	0.63	0.59
4. During important tests I am so tense that my stomach gets upset.	0.689	0.423	0.71	0.97	1.07
5. I seem to defeat myself while working on important tests.	0.425	0.651	-0.14	1.19	1.18
6. I feel very panicky when I take an important test.	0.781	0.521	0.14	0.69	0.71
7. I worry a great deal before taking an important examination.	0.744	0.394	- 0.59	0.76	0.78
8. During tests I find myself thinking about the consequences of failing.	0.556	0.668	0.07	0.99	0.98
9. After an exam is over I try to stop worrying about it but I can't.	0.558	0.494	0.65	1.02	1.02
10. During examinations I get so nervous that I forget facts I really know.	0.585	0.628	0.04	0.91	0.97

Short TAI calibrations

The Rasch model was used for the calibrations. The first calibration on the full dataset revealed no misfitting items therefore no other calibration was considered necessary.

Item-person maps are presented to show how well the items are targeted for the population of students and finally the students are divided into groups according to their test anxiety estimates for investigating later on whether test anxiety is associated with misfit.

The first calibration of the full set of data included 504 students (12 students, 5 maximum scorers and 7 minimum scorers were removed from the calibration) and 10 items.

Table 4.2.31 below shows the item statistics in misfit order.

Table 4.2.31 Items statistics: misfit order

ENTRY	RAW				INFIT		OUTFIT		PTMEA	
NUMBER	SCORE	COUNT	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	ZSTD	CORR.	items
9	981	492	.54	.07	1.14	2.2	1.23	3.0	A .63	item 9
8	1027	492	.34	.07	1.21	3.2	1.16	2.3	B .66	item 8
4	933	492	.76	.07	1.18	2.6	1.05	.7	C .69	item 4
2	1042	492	.28	.06	1.17	2.6	1.10	1.4	D .68	item 2
5	1167	492	-.23	.06	1.09	1.4	1.16	2.4	E .63	item 5
10	1158	492	-.19	.06	1.00	.0	.97	-.4	e .72	item 10
1	1214	492	-.41	.06	.84	-2.7	.84	-2.7	d .71	item 1
7	1339	492	-.91	.06	.84	-2.8	.83	-2.7	c .74	item 7
3	1189	492	-.32	.06	.82	-3.2	.79	-3.4	b .74	item 3
6	1077	492	.13	.06	.80	-3.5	.76	-4.0	a .76	item 6
MEAN	1113.	492.	.00	.06	1.01	.0	.99	-.3		
S.D.	116.	0.	.48	.00	.16	2.6	.17	2.5		

There are no misfitting items; therefore the scale data fit the Rasch model very well. All the items have infit or outfit much smaller than the cut-off score of 1.5 used in the analyses of scales in this study. Given the good fit of the items to the model, there was no need for a second calibration.

A summary of the results of the Rasch analysis from the calibration is given in table 4.2.32.

Table 4.2.32 Summary of the results of the Rasch analysis for the TAI

	N	Estimate of mean (SD)	Range	Reliab.	Separ. Index	Infit msq mean (SD)	Outfit msq mean (SD)
Examinees	504	- 0.45 (1.38)	-3.83 - 3.61	0.85	2.42	0.99 (0.53)	0.99 (0.53)
Items	10	0.0 (0.48)	-0.91 - 0.76	0.98	7.03	1.01 (0.16)	0.99 (0.17)

The word ‘ability’ is used in the place of ‘test anxiety measure’ in order to be consistent with analyses of the other tests.

The range of student abilities was from -3.83 to 3.61 (excluding the maximum and minimum scorers whose estimates were 4.84 and – 5.10), with a mean of -0.45 (SD = 1.38). The reliability of student estimates was 0.85. This index is equivalent to Cronbach’s alpha (alpha = 0.89). The student separation index was 2.42. This indicates the spread of person measures in standard error units; in this case it is just over 2.4 standard errors. The higher the value of this index, the more spread out the persons are on the variable being measured.

A student separation index of 2.42 also indicates approximately just over 3.5 statistically distinct strata (strata = 3.56) of student abilities identified by the instrument,

The item estimates ranged from – 0.91 to 0.76 and the reliability index was 0.98. 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 7.03, indicating that the spread of item estimates is about 7 standard errors.

There were only 3 badly misfitting students (outfit and/or infit > 3.0), however they were not removed since the data already fitted the Rasch model well.

Figure 4.2.19 shows the item-student map. The items seem to be well targeted for students with abilities around the mean ability. The item measures lie between half a standard deviation below and 1 standard deviation above the overall mean student ability.

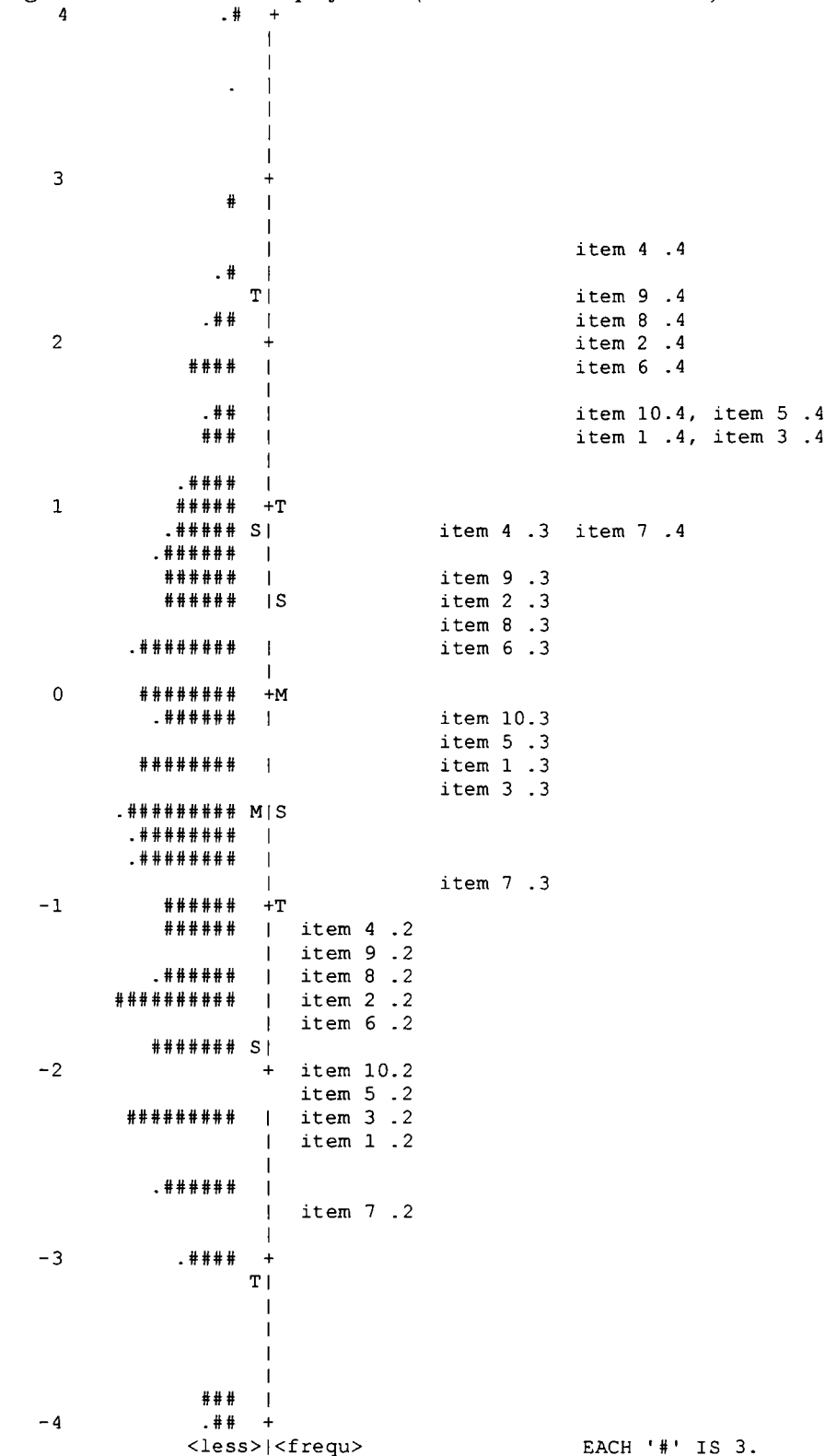
Figure 4.2.19 Students map of items



Figure 4.2.20 shows another item – student map, this time with all the categories of the items (the thresholds for all the possible scores for each item). It is obvious that the various steps of the questions are well targeted for a much wider range of abilities.

That is, from about 2 standard deviations below to just above 2 standard deviations of the overall mean ability.

Figure 4.2.20 students map of items (with item score thresholds)



4.2.10 Reliability and Validity of the TAI

For the validation study of the TAI, the following evidence was collected and presented below:

- *Comparisons of short TAI results with TAI results from the first phase*
- *Principal components analysis of the raw scores*
- *Correlations of the short TAI scores with maths test scores*
- *Principal components analysis of the standardised residuals (based on Rasch analyses)*

Reliability

The reliability of student estimates was 0.85 and Cronbach's alpha 0.89. These are considered high values for questionnaires.

Validity

Comparisons of short TAI results with TAI results from the first phase

Cronbach's alpha for the TAI was 0.924 and for the short TAI 0.895. The reason for this difference is the length of the scale. The original TAI consisted of 20 items whereas the shorter version of 10 items.

Using the Spearman-Brown formula on the short TAI reliability, for estimating the reliability coefficient for an instrument with double the length, gives 0.944. This estimate is slightly higher than the reliability of 0.924 and this can easily be explained since the items used for the shorter version of the instrument were the ones with the higher loadings on the two factors measured.

Furthermore, in both scales the female students had significantly higher scores than the male students, as shown in table 4.2.33.

Table 4.2.33 Comparisons of male-female scores on the two TAIs

Gender	Mean score on TAI	p-value	Mean score on short TAI	p-value
Male	42.36	0.000	20.76	0.000
Female	47.77		23.88	

Principal components analysis

Principal components analysis extracted only one factor which ‘explains’ about 52% of the variation in the data. Table 4.2.34 shows the factor extracted.

Table 4.2.34 Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5,156	51,556	51,556	5,156	51,556	51,556
2	,968	9,678	61,234			
3	,669	6,688	67,922			
4	,574	5,740	73,663			
5	,527	5,272	78,935			
6	,480	4,802	83,737			
7	,452	4,522	88,259			
8	,417	4,174	92,433			
9	,411	4,108	96,541			
10	,346	3,459	100,000			

Extraction Method: Principal Component Analysis.

There was only one factor extracted in the short TAI and all the items loaded highly significantly (loadings 0.641 to 0.793), as shown in table 4.2.33 because:

- The items with the highest correlation with the two factors from the original TAI were selected and used
- The loadings of the items were significant on both factors in the original analyses and
- The two factors were significantly correlated ($r = 0.636$).

The researcher was not interested in breaking up test anxiety into the two factors but in measuring the students’ test anxiety with a shorter, and easier to administer, questionnaire.

Table 4.2.35 shows the items used in this shorter version of the TAI and the loading of each item on the factor extracted.

Table 4.2.35 Loadings of the items

Statement	Factor
1. Thoughts of doing poorly interfere with my concentration on tests.	0.729
2. I feel very jittery when taking an important test.	0.707
3. During tests I feel very tense.	0.759
4. During important tests I am so tense that my stomach gets upset.	0.726
5. I seem to defeat myself while working on important tests.	0.633
6. I feel very panicky when I take an important test.	0.793
7. I worry a great deal before taking an important examination.	0.766
8. During tests I find myself thinking about the consequences of failing.	0.672
9. After an exam is over I try to stop worrying about it but I can't.	0.641
10. During examinations I get so nervous that I forget facts I really know.	0.736

Correlations with maths test scores.

Table 4.2.36 shows the correlations of the short TAI scores with the scores on diagnostic test, the test on the quadratic equations and the final maths exam.

Table 4.2.36 Correlation with maths tests

	Diagnostic test	Maths test	Maths exam
Short TAI	- 0.300	- 0.309	- 0.314
N	504	383	496

All the correlations were significant ($p < 0.01$).

The above analyses (Principal components analysis and correlations) were performed to simply reconfirm what the Rasch analyses show (Rasch validation study is presented next).

Principal components analysis of the standardised residuals

Principal components analysis on the standardised residuals (Linacre, 1988) was performed in WINSTEPS yielding:

PRINCIPAL COMPONENTS (STANDARDIZED RESIDUAL) FACTOR PLOT				
Factor 1 extracts 1.9 units out of 10 units of item residual variance noise.				
Yardstick (variance explained by measures)-to-This Factor ratio: 11.1:1				
Yardstick-to-Total Noise ratio (total variance of residuals): 2.1:1				
Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)				
		Empirical		Modeled
Total variance in observations	=	31.2	100.0%	100.0%
Variance explained by measures	=	21.2	68.0%	67.3%
Unexplained variance (total)	=	10.0	32.0%	32.7%
Unexpl var explained by 1st factor	=	1.9	6.1%	

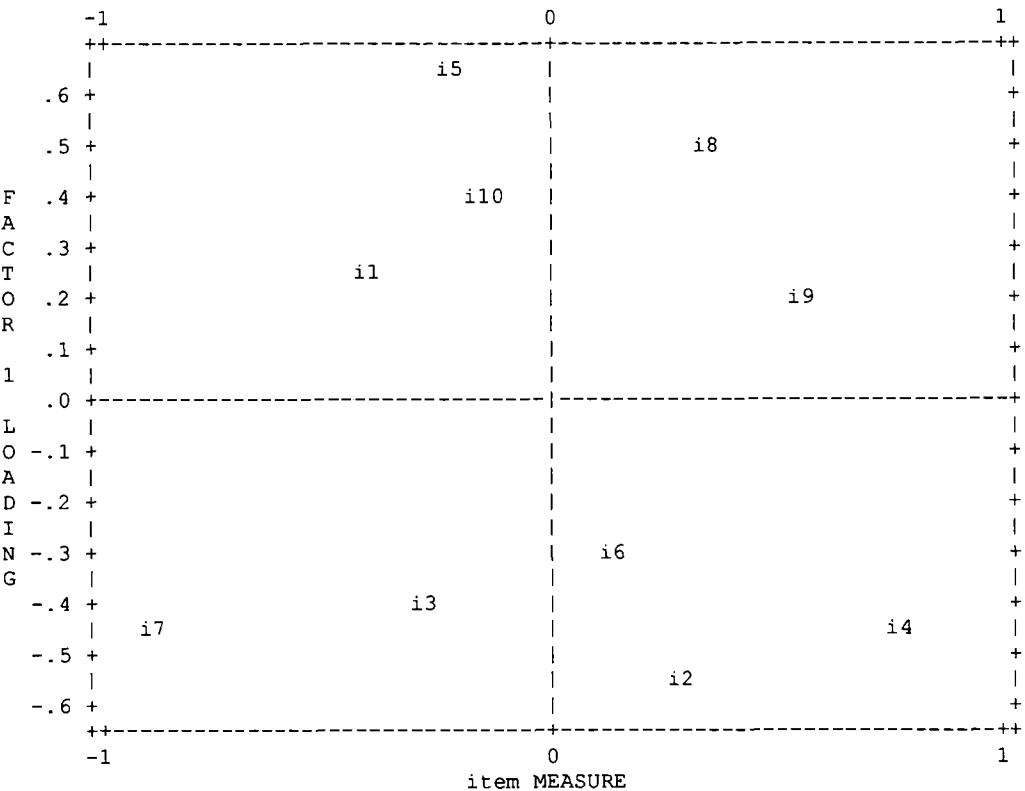
The variance explained by the measures (i.e. by the dimension measured by the scale) is 68 % of the total variance. It is also more than 11 times the variance explained by the first factor extracted by PCA on the standardised residuals and more than 2 times the total unexplained variance in the data. The unexplained variance is 32 % of the total variance in the data.

Also, the variance explained by this first factor is 19 % of the unexplained variance (1.9 out of 10) and just 6.1 % of the total variance in the data.

These figures support the unidimensional structure of the data.

Figure 4.2.21 shows the plot of the loadings of the items on the first factor extracted against their measures.

Figure 4.2.21 Factor loadings against item measures.



Both methods (PCA of the raw scores and PCA of the standardised residuals) agree on the fact that there is no second dimension present in the data.

However, the plot of factor loadings against item measures (figure 4.2.16) separates the items, based on their loadings on the first factor extracted after removing the dimension measured by the scale, into two groups. The top group contains the items shown to be measuring the worry factor (and item 9 measuring total anxiety) on the original 20-item TAI and the bottom group the items measuring the emotionality factor (and item 4 measuring total anxiety).

The separation of the items based on the two factors, which combined make the test anxiety dimension, is very useful diagnostically since it shows where the items originated from.

It by no means suggests a second dimension in the data.

The unidimensional structure in the Rasch approach is supported by:

- The good fit of the data to the Rasch model
- The numbers and percentages reported in the PCA of the standardised residuals

- The highly significant correlation ($r = 0.71$) between the total scores on the two item groupings shown in the figure above.

Different anxiety groups

The range of abilities (anxiety measures) was divided into three different groups, the low, medium and top anxiety groups using three different cut-off scores, the 30th and 70th percentiles, the 20th and 80th percentiles and the 10th and 90th percentiles.

4.2.11 Misfitting students

Misfitting students were identified using appropriate cut-off scores for the infit and outfit statistics (1.5 for both). The numbers and proportions of misfitting students are presented, together with comparisons of equivalent proportions from a simulation study.

Then, an investigation was carried out into whether the same students misfit in administrations of different psychometric scales.

Following the calibration of the TAI, misfitting students were identified using cut-off scores for the infit and outfit mean squares of 1.5.

Table 4.2.37 shows the number of students identified as misfitting by the two indices as well as the total number.

Table 4.2.37 Misfit (infit) * Misfit (outfit) Crosstabulation

		Misfit (outfit)		Total
		Fitting	Misfitting	
Misfit (infit)	Fitting	425	8	433
	Misfitting	5	66	71
Total		430	74	504

The number of students identified as misfitting by the infit statistic was 71 (14.1%) and by the outfit statistic was 74 (14.7%) whereas 66 students were identified by both, giving a total of 79 (15.7%) misfitting students.

Do the same students misfit in different administrations of psychometric scales?

Table 4.2.38 shows the numbers of fitting and misfitting students in the Self-esteem scale (MSES) and the short TAI.

It shows that 14.6% (55 out of 377) of the fitting students in the MSES and 19.0% (16 out of 84) of the misfitting students in the MSES were also misfitting in TAI.

Similarly, 17.4% (68 out of 390) of the fitting students in TAI and 22.5% (16 out of 71) of the misfitting students in TAI were also misfitting in the diagnostic test.

The Chi-square test performed (for association between misfit in the TAI and misfit in the MSES) yielded a chi-square statistic of 0.734 ($p = 0.392$) and a non-significant result. That is, there is no association between misfittings in the two scales.

Table 4.2.38 Misfit in TAI * Misfit in MSES Crosstabulation

		Misfit in TAI		Total
		Fitting Students	Misfitting Students	
Misfit in MSES	Fitting Students	322	55	377
	Misfitting Students	68	16	84
Total		390	71	461

Chi-square = 0.734, d.f. = 1, $p = 0.392$

4.2.12 Detecting multidimensionality: PCA of Rasch standardised residuals or of raw score?

The researcher had the responses of 298 students to 27 maths items (from the maths diagnostic test) and to 28 language items (from the language diagnostic test). All the data were put together, as a 55-item test, and were analysed using first PCA of the raw scores followed by Rasch analyses and PCA of the standardised residuals.

PCA of the raw scores

Table 4.2.39 below shows the factors extracted with PCA of the raw scores.

Table 4.2.39 Total Variance Explained (Extraction Method: Principal Component Analysis)

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	11,919	21,671	21,671	11,919	21,671	21,671
2	3,002	5,458	27,129	3,002	5,458	27,129
3	2,104	3,825	30,955	2,104	3,825	30,955
4	1,916	3,483	34,438	1,916	3,483	34,438
5	1,801	3,275	37,713	1,801	3,275	37,713
6	1,564	2,844	40,557	1,564	2,844	40,557
7	1,474	2,679	43,236	1,474	2,679	43,236
8	1,387	2,521	45,757	1,387	2,521	45,757
9	1,309	2,380	48,138	1,309	2,380	48,138
10	1,261	2,292	50,430	1,261	2,292	50,430
11	1,219	2,216	52,646	1,219	2,216	52,646
12	1,180	2,146	54,793	1,180	2,146	54,793
13	1,136	2,066	56,858	1,136	2,066	56,858
14	1,083	1,970	58,828	1,083	1,970	58,828
15	1,053	1,914	60,742	1,053	1,914	60,742
16	1,004	1,826	62,567	1,004	1,826	62,567

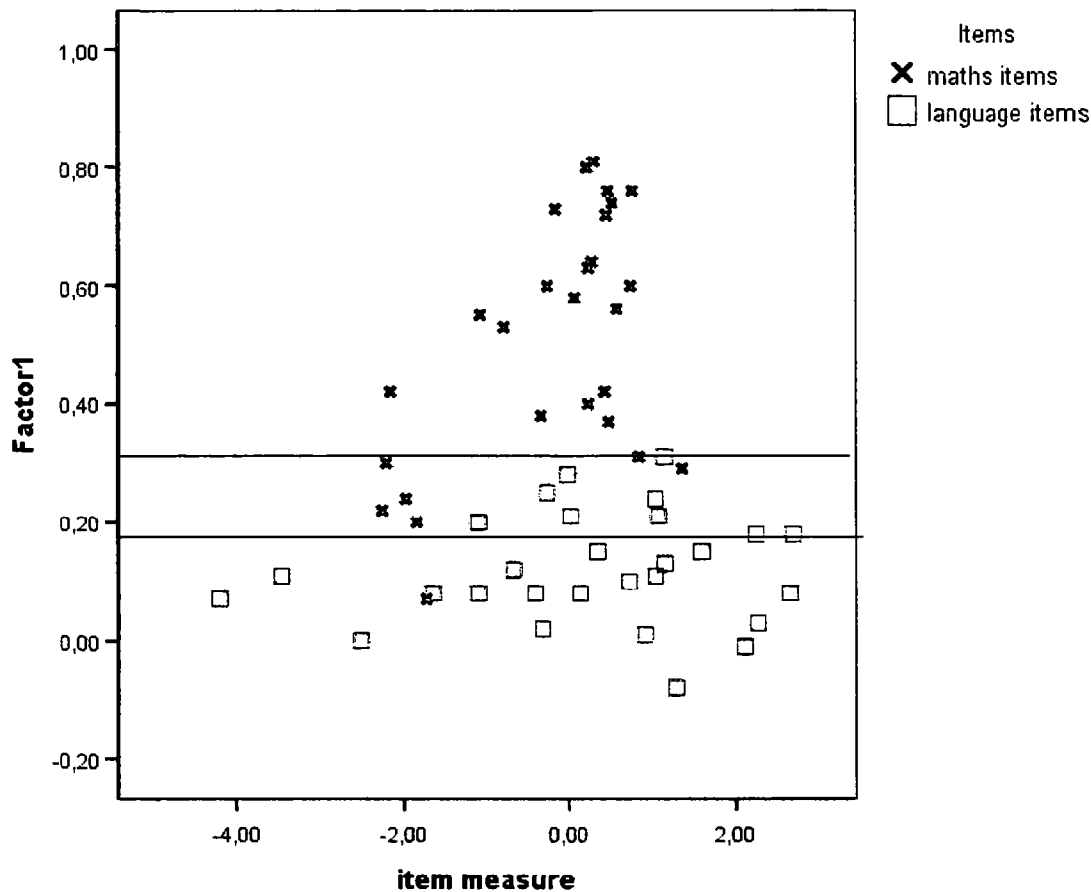
PCA extracted 16 factors (eigenvalue > 1).

The eigenvalue grand total is 55, the number of items. The problem is how many factors to report. According to Linacre (personal communication, December 9, 2007), simulation studies indicate that loadings of less than 1.4 can happen by chance. In practice however, we are only concerned with factors more than 2 or more items worth of information, in this case the first three factors in table 4.2.39.

Following this, the researcher plotted the loadings of the items on the first three factors, in order to investigate the dimensionality of the test.

The following figures, 4.2.21, 4.2.22 and 4.2.23 show scatter plots of the factor loadings against item measure (in order to make the plots comparable with the ones in Rasch analyses) for the first three factors extracted by the PCA of the raw scores.

Figure 4.2.21 Factor 1 loadings against item measure



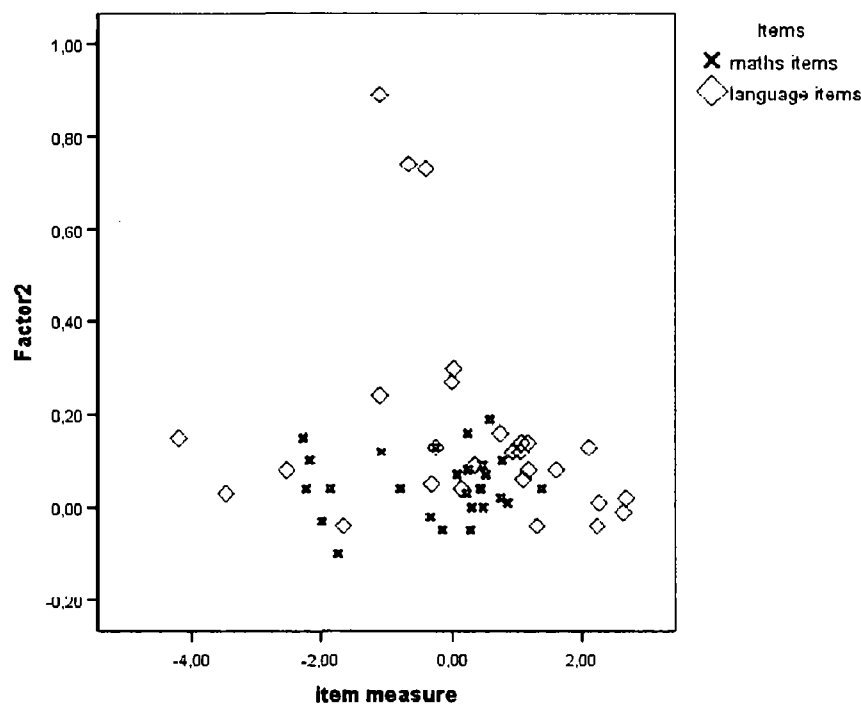
There is a good separation of the maths from the language items, with the maths items having in general higher loadings on factor 1. However, there is a range (from about 0.18 to 0.31) where there are 6 maths items and 9 language items.

This factor has an eigenvalue of 11.9, that is, the strength of about 12 items. These are roughly the 12 items with the big loadings in the plot. These items are all maths items therefore one can conclude that factor 1 can be interpreted as maths ability.

In the next two figures, the loadings on factors 2 and 3 reveal no separation of the two dimensions.

Factor 2 has eigenvalue 3.002 that is, the strength of 3 items; the 3 items with the highest loadings (shown as outliers in figure 4.2.22) on factor 2 are language items. This factor can be interpreted, in a similar way as before, as Language ability.

Figure 4.2.22 Factor 2 loadings against item measure



There is no obvious separation of the maths items from the language items, except from the 3 language items which look like outliers and load very highly with this factor.

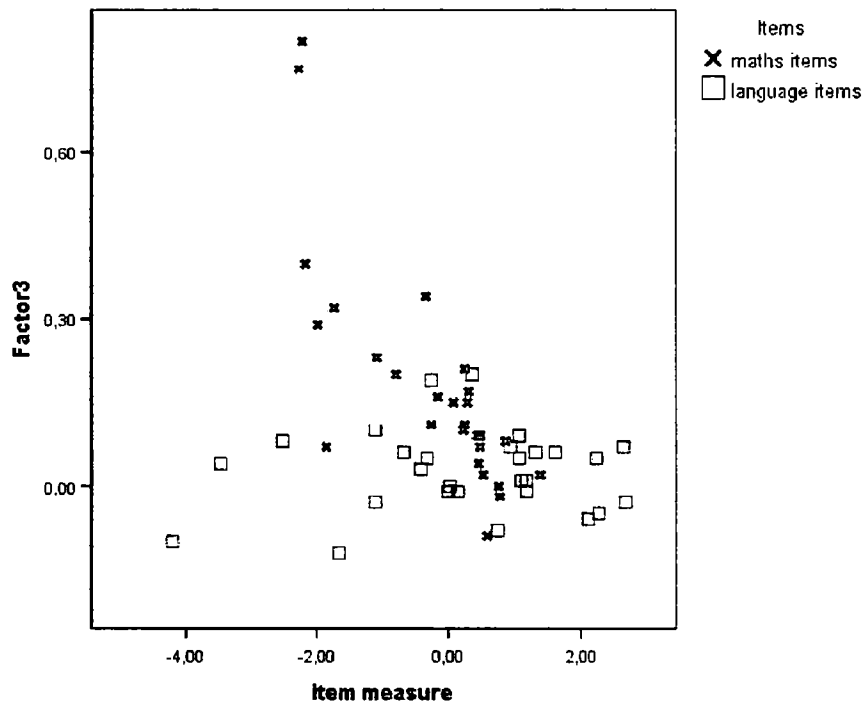
Factor 3 has eigenvalue 2.104 that is, the strength of about 2 items; the 2 items with the highest loadings (shown as outliers in figure 4.2.23) on factor 3 are maths items.

The two items are very simple algebraic items. They are asking students to complete the following:

$x + x = \dots\dots\dots$ and $x \cdot x = \dots\dots\dots$

This makes the interpretation rather difficult; one could perhaps say that this factor describes an understanding of very simple algebraic calculations.

Figure 4.2.23 Factor 3 loadings against item measure



There is again no obvious separation of the maths items from the language items, except from the 2 maths items which look like outliers and load very highly with this factor and a cluster of 4 items just above bulk of the points.

These analyses have focused on only the first 3, and more significant, factors of the PCA.

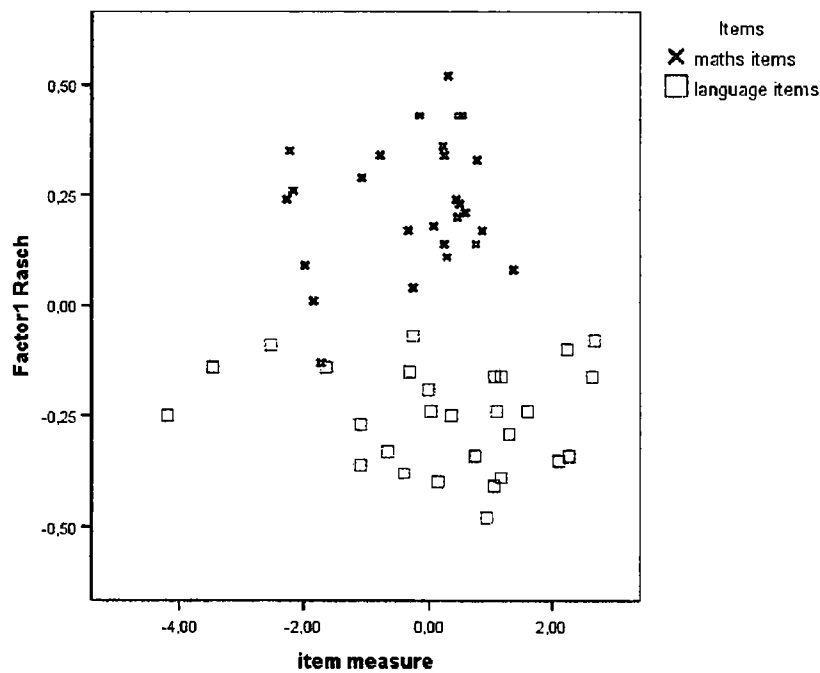
PCA of the Rasch standardised residuals

The data were analysed using the Partial Credit Rasch model and an investigation of the dimensionality was carried out through PCA of the standardised residual, giving:

PRINCIPAL COMPONENTS (STANDARDIZED RESIDUAL) FACTOR PLOT				
Factor 1 extracts 4.1 units out of 55 units of item residual variance noise.				
Yardstick (variance explained by measures)-to-This Factor ratio: 46.0:1				
Yardstick-to-Total Noise ratio (total variance of residuals): 3.4:1				
Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)				
		Empirical		Modeled
Total variance in observations	=	243.7	100.0%	100.0%
Variance explained by measures	=	188.7	77.4%	78.0%
Unexplained variance (total)	=	55.0	22.6%	22.0%
Unexpl var explained by 1st factor	=	4.1	1.7%	

Figure 4.2.24 shows the factor 1 loadings against item measure

Figure 4.2.24 Factor 1 against item measure



Factor 1 plots from the two approaches are more or less telling the whole story, however the context has changed.

In the PCA of the raw scores, the factor 1 plot includes the correlation with the latent variable; therefore almost all loadings are positive. Thus in that plot we cannot see how big the maths vs language effect is because it is combined with the maths + language vs latent variable effect.

In the PCA of the standardised residuals however, the latent variable is excluded, so the loadings are balanced around 0. The PCA of standardised residuals shows the maths vs language effect in Factor 1 which has an eigenvalue of 4.1.

In the PCA of raw scores factor 1 has eigenvalue of approximately 12, composed of about 4 of maths vs language effect and 8 of maths + language on the latent variable. From that, the latent variable (8) looks only twice the strength of the maths vs language factor (4), and that suggests multidimensionality.

But the PCA of standardised residuals shows that the maths + language on the latent variable (variance explained by the measures) is very strong. It is 188.7 (77.4% of the total variance). This is because the raw scores PCA has lost the variance in the data explained by differences in person measures and item measures and retained only the differences in the data explained by the correlations.

So is the data unidimensional? The maths vs language effect has a strength of 4.1 which explains only about 7.5% of the unexplained variance and only 1.7% of the total variance in the residuals. According to Linacre (personal communication, December 9, 2007), the split shown in the factor 1 plot may be useful diagnostically in the classroom. (If the two sets of items were taken separately then indeed they would measure two different constructs, that is, maths and language abilities)

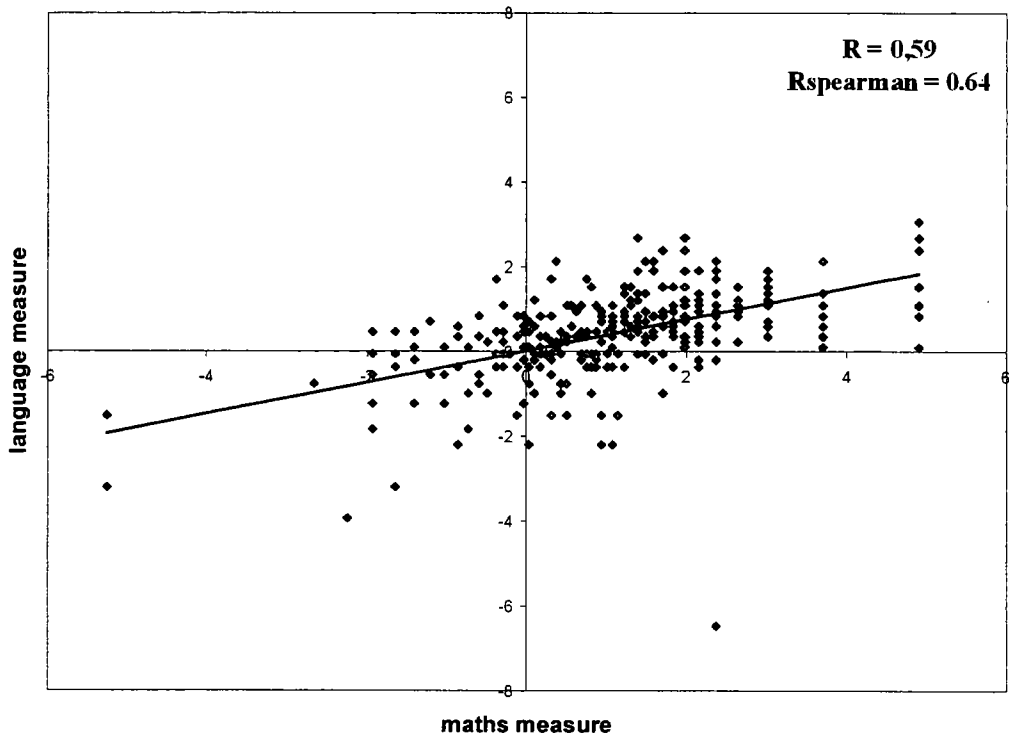
But for policy-makers the language and math items are telling the same story. Therefore, the data are unidimensional.

Comparing PCA of raw scores with PCA of standardised residuals one can say that, the interpretation of PCA of raw scores is usually very difficult and so is the decision as to whether the test is unidimensional, or if it is not, as to what the dimensions are. (Confirmatory analysis may be helpful in deciding about the unidimensionality in unclear cases)

Furthermore, very wide and very narrow ranges of person and item measures can produce the same correlation matrices and so the same eigenvalues, but the Rasch measures would explain very different amounts of the total variance on the observations depending on the ranges of person and item measures. The wider the range of the measures the more variance the measures explain in the observations.

The researcher thought that one final investigation into the unidimensionality of the test, as decided through the Rasch model, would be to estimate the students' measures separately for the maths items and the language items and plot them on a scatter diagram. If there is a strong relationship between the two then that will support the conclusion of unidimensionality. Figure 4.2.25 shows the plot of the persons' measures from the language and the maths items.

Figure 4.2.25 Language measures vs maths measures



There is an obvious outlier in the diagram, the lowest point in the fourth quadrant, which when removed the correlation coefficient is 0.59 and the Spearman Rank correlation coefficient is 0.64. These are significant correlations supporting the conclusion of a unidimensional structure.

It is the researcher's belief that the fact that both tests used were diagnostic and targeted for the lower ability students (i.e. easy tests) contributes to the unidimensional structure of the data, since it was not difficult for many students to perform well on both sets of items (most points are in the first quadrant of the figure. In other words one could hypothesise that the test measures a general academic ability and may be weighted towards general intelligence (g, derived by Spearman in 1904).

4.2.13 Investigation of possible factors associated with misfit

Log-linear analysis was performed in an attempt to investigate possible association of various factors with misfit separately in the two maths tests. The saturated models considered included:

$$\begin{aligned} & \text{Student gender} * \text{Misfit} \\ & \text{Student gender} * \text{Ability} * \text{Test Anxiety} * \text{Misfit} \\ & \text{Student gender} * \text{Ability} * \text{Maths Self-esteem} * \text{Misfit} \end{aligned}$$

Misfit in Test 1 (the diagnostic test)

First possible reasons for aberrance were investigated using the fitting and misfitting students in the first test, the one with the multistep problems, items with the same format as the test items in phase 1.

The tables below show the saturated model used in each case and the significant effects, if any, based on the partial associations derived from the Likelihood-ratio chi-square.

$$\text{Student gender} * \text{Misfit}$$

No significant association was found between student gender and misfit in test 1.

$$\text{Student gender} * \text{Ability} * \text{Test Anxiety} * \text{Misfit}$$

Table 4.2.40 shows the results of the analysis when Student gender * Ability * Test Anxiety * Misfit was used. The categorical variables Ability and Test Anxiety with cut-off scores at the 30th and 70th percentiles were used.

Table 4.2.40 Partial Associations of significant association or interaction terms

Saturated model: Student gender * Ability * Test Anxiety * Misfit			
Two-way effects	L ²	d.f.	p-value
Students gender * Test Anxiety	10.435	2	0.005
Ability * Test Anxiety	18.480	4	0.001

No association was found between student gender, ability, test anxiety, or any combination of those variables, with misfit.

The association between student gender and test anxiety was, as expected, significant, just as in phase 1. This was also evident in the comparison between the mean scores of male and female students on the test anxiety scale, with the females scoring significantly higher.

Similarly, given the significant negative correlation between test anxiety and scores on the test, the significant association between ability and test anxiety was expected too.

When the 20th, 80th, and the 10th, 90th percentiles were used as the cut-off scores for the test anxiety and the ability categorical variables no significant associations were found between any of the variables, apart from ability and anxiety at the 10th and 90th percentiles ($p = 0.009$).

*Student gender * Ability * Maths Self-esteem * Misfit*

Table 4.2.41 shows the results of the analysis when the model: Student gender * Ability * Maths Self-esteem * Misfit was used. The categorical variables Ability and Maths Self-esteem with cut-off scores at the 30th and 70th percentiles were used.

Table 4.2.41 Partial Associations of significant association or interaction terms

Saturated model: Student gender * Ability * Maths Self-esteem * Misfit			
Two-way effects	L²	d.f.	p-value
Ability * Self-esteem	73.664	4	0.000

No association was found between student gender, ability, Maths Self-esteem, or any combination of those variables, with misfit.

The only association found was between Ability and Maths Self-esteem which again was expected given the significant positive correlation found between those two continuous variables (before they were transformed into categorical variables)

Identical results (no associations except from Ability with Maths Self-esteem) were found when the 20th, 80th, or the 10th, 90th, percentiles were used as cut-off scores for the variables of Ability and Maths Self-esteem.

Misfit in Test 2

The main difference between this test and the other two maths tests used throughout this study was the fact that the majority of the items (12 out of 16) were dichotomous multiple-choice items.

The tables below again show the saturated model used in each case and the significant effects, if any, based on the partial associations derived from the Likelihood-ratio chi-square.

*Student gender * Misfit*

No significant association was found between student gender and misfit in test 1.

*Student gender * Ability * Test Anxiety * Misfit*

Table 4.2.42 shows the results of the analysis when Student gender * Ability * Test Anxiety * Misfit was used. The categorical variables Ability and Test Anxiety with cut-off scores at the 30th and 70th percentiles were used.

Table 4.2.42 Partial Associations of significant association or interaction terms

Saturated model Student gender * Ability * Test Anxiety * Misfit				
Effect	Model	L ²	d.f.	p-value
3 – way	Gender * Anxiety * Misfit	7.598	2	0.022
2 – way	Ability * Misfit	8.084	2	0.018

Significant associations were found between ability and misfit and between the interaction of gender with anxiety on misfit.

Further investigation was undertaken into the effect of ability on misfit and the interaction of gender with anxiety on misfit.

Table 4.2.43 shows the crosstabulation of misfit * ability with the proportions of fitting and misfitting students in the 3 different levels of ability used.

Table 4.2.43 Ability * Misfit crosstabulation

		Misfit		Total
		Fitting	Misfitting	
Ability level	Low 30%	104 (74.3%)	36 (25.7%)	140
	Medium	115 (71.0%)	47 (29.0%)	162
	Top 30%	89 (62.2%)	54 (37.8%)	143
Total		308	137	445

The proportion of misfitting students among the top ability students is significantly higher than the proportion among the other two categories ($p = 0.018$).

Table 4.2.44 shows the Gender * Test Anxiety * Misfit crosstabulation with the proportions of fitting and misfitting students in the 3 different levels of ability used, separately for male and female students.

Table 4.2.44 Gender * Test Anxiety * Misfit crosstabulation

Gender	Test Anxiety level	Misfit		Total
		Fitting	Misfitting	
Male	Low 30%	45 (73.8%)	16 (26.2%)	61
	Medium	32 (58.2%)	23 (41.8%)	55
	Top 30%	25 (80.6%)	6 (19.4%)	31
Total		102	45	147
Female	Low 30%	28 (54.9%)	23 (45.1%)	51
	Medium	71 (71.0%)	29 (29.0%)	100
	Top 30%	60 (70.6%)	25 (29.4%)	85
Total		159	77	239

The above table shows that the association between test anxiety and misfit has a significantly different pattern for male students than for female students.

Figure 4.2.26 below shows the pattern for the males. The medium anxiety group has the highest proportion of misfitting students and the top anxiety group the lowest proportion of misfitting students.

Figure 4.2.26 Misfit at different anxiety levels in Males

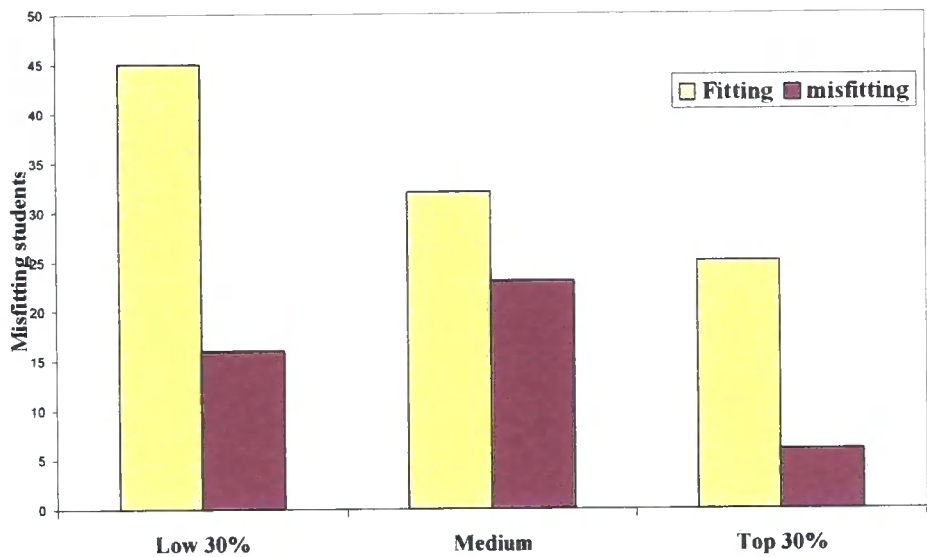
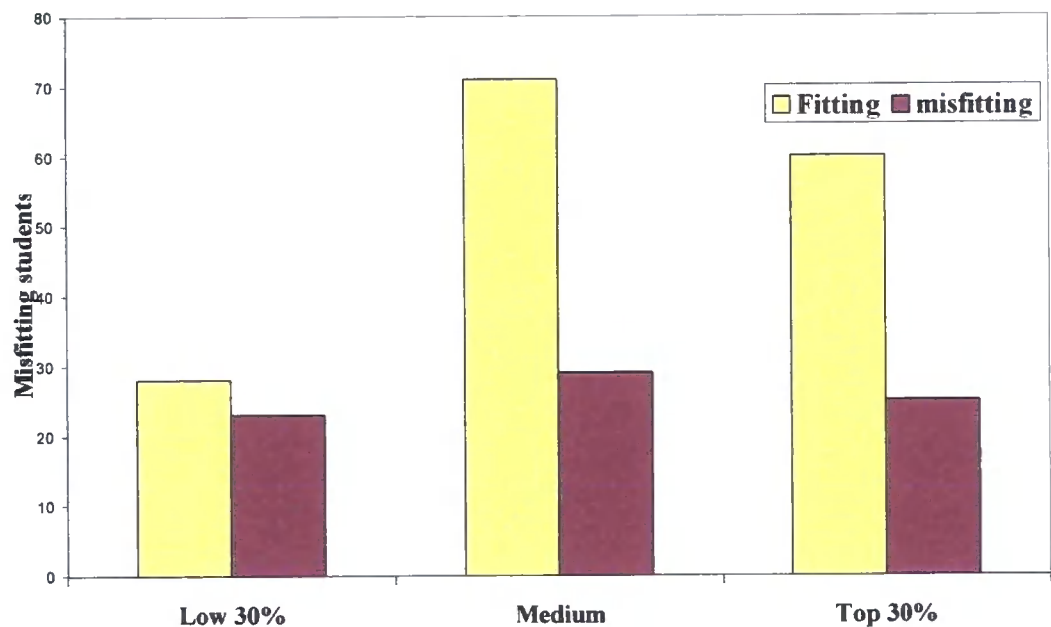


Figure 4.2.27 shows the pattern for the females. The highest proportion of misfitting students is in the low anxiety group and the other two anxiety groups (medium and top) have the same proportions.

Figure 4.2.27 Misfit at different anxiety levels in Females



No significant associations were found when the 20th, 80th, or the 10th, 90th, percentiles were used as cut-off scores for the variables of Ability and Test Anxiety except from the interaction of ability and anxiety on misfit in the case of the 10th and 90th percentiles. ($L^2 = 10.345$, d.f. = 4 and $p = 0.035$).

*Student gender * Ability * Maths Self-esteem * Misfit*

Table 4.2.45 shows the results of the analysis when the model: Student gender * Ability * Maths Self-esteem * Misfit was used. The categorical variables Ability and Maths Self-esteem with cut-off scores at the 30th and 70th percentiles were used.

Table 4.2.45 Student gender * Ability * Maths Self-esteem * Misfit

Saturated model: Student gender * Ability * Maths Self-esteem * Misfit			
Two-way effects	L ²	d.f.	p-value
Gender * Ability	8.021	2	0.018
Ability * Self-esteem	25.838	4	0.000

No association was found between student gender, ability, Maths Self-esteem, or any combination of those variables, with misfit.

The significant association between Ability and Maths self-esteem was expected because of the highly significant correlation (0.617) between the scores on the MSES and the test 2.

A significant association was found between gender and ability since more females (36.3%) were categorised in the top 30% ability group than the males (26.3%).

No significant associations were found when the 20th, 80th, or the 10th, 90th, percentiles were used as cut-off scores for the variables of Ability and Maths self-esteem except from the association of ability and misfit in the case of the 10th and 90th percentiles. ($L^2 = 6.057$, d.f. = 2 and $p = 0.048$).

4.2.14 Predictive validity of the test scores of fitting and misfitting students

The hypothesis under investigation was that the predictive validity of the scores of misfitting students is of a lower degree than of the fitting students. For the purposes of this investigation the test scores of fitting and misfitting students in both tests (done separately) were correlated with other criteria. 95% confidence intervals were calculated and comparisons were made. The other criteria used were: a second maths test, the first term maths grade and the scores in the final maths exam.

Maths Test 1 (the Diagnostic test)

The test scores of fitting and misfitting students in the first test (the diagnostic) were correlated with other criteria to investigate whether there are any differences in their predictive validity. The hypothesis was that the predictive validity of the misfitting students is of a lower degree than that of the fitting students, thus the correlation coefficients are lower. The other criteria involved the first term grade in maths, the scores in a second maths test and the scores in the final maths exam.

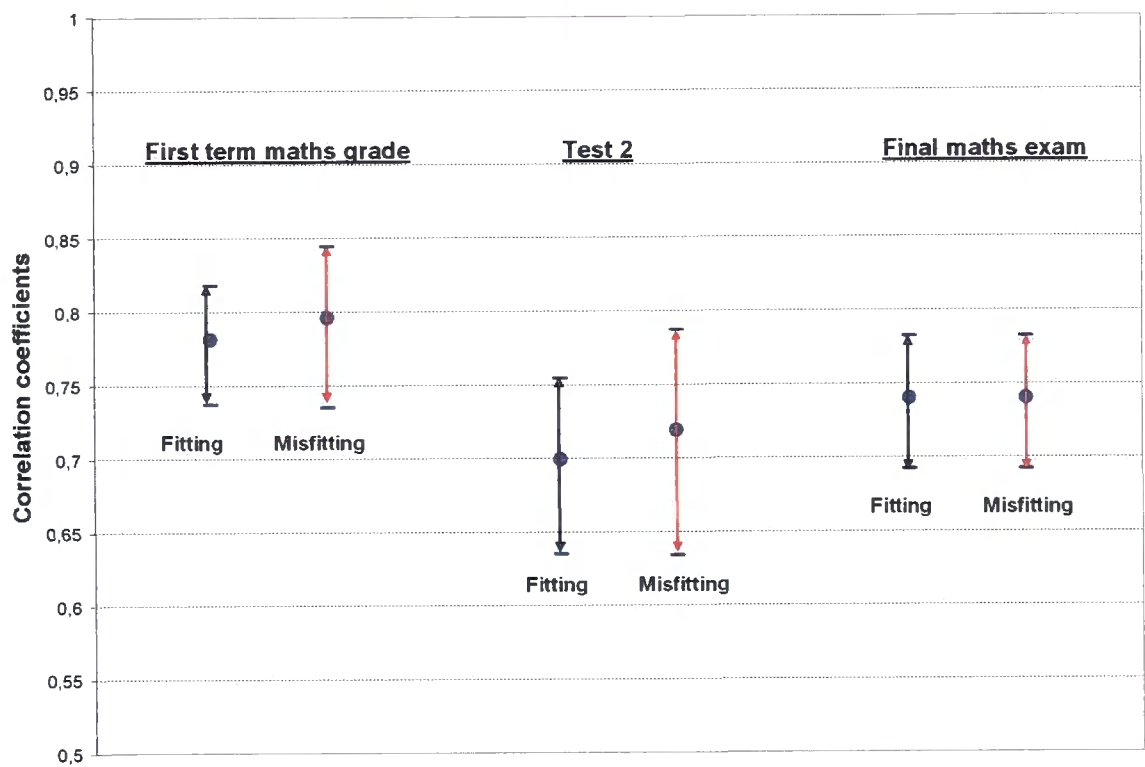
Table 4.2.46 shows the correlation coefficients of the scores of fitting and misfitting students in test 1 with the other criteria. It also shows 95% confidence intervals for each coefficient.

Table 4.2.46 Correlations and 95% C.I. for test 1 scores with other criteria for fitting and misfitting students

	1 st term maths grade		Test 2		Final maths exam	
	Fitting	Misfitting	Fitting	Misfitting	Fitting	Misfitting
Correlation coeff. (N)	0.781 (356)	0.796 (177)	0.699 (288)	0.719 (157)	0.740 (395)	0.740 (210)
Upper (95%) limit	0.818	0.844	0.754	0.787	0.782	0.796
Lower (95% limit)	0.737	0.735	0.635	0.634	0.692	0.672

Figure 4.2.28 shows the correlation coefficients and the 95% confidence intervals in a diagrammatic form, to make comparisons easier.

Figure 4.2.28 Correlations and 95% C.I. for test 1 scores with other criteria



There is absolutely no evidence of a difference in the correlations of the test scores of fitting and misfitting students with the other criteria.

The standard deviations of the scores of fitting and misfitting students were very similar: For test 1 they were 13.84 and 12.82, for the first term grade 3.52 and 3.52, for test 2 they were 6.36 and 6.49 and for the final exam 6.10 and 6.02 respectively.

Test 2

The test scores of fitting and misfitting students in test 2 were correlated with other criteria to investigate whether there are any differences in their predictive validity. The hypothesis was again that the predictive validity of the misfitting students is of a lower degree than that of the fitting students, thus the correlation coefficients are lower. The important difference between the two tests was the item format, with the second test containing 12 (out of a total of 16) multiple-choice items. The other criteria involved the

first term grade in maths, the scores in first maths test and the scores in the final maths exam.

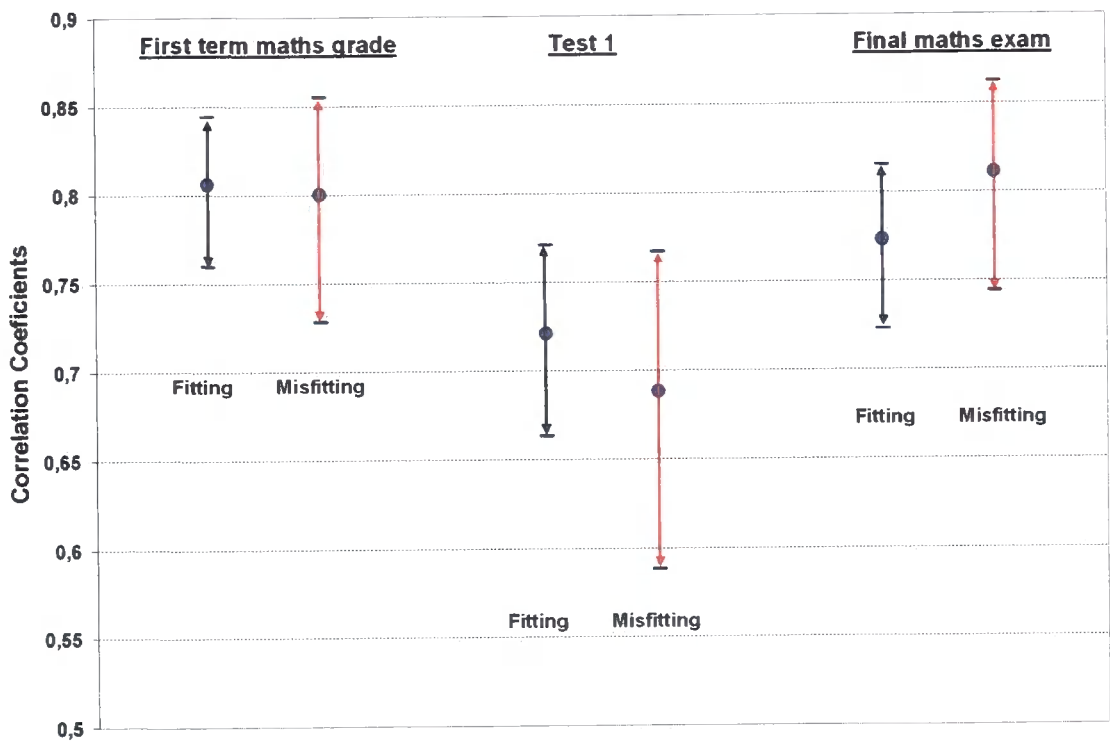
Table 4.2.47 shows the correlation coefficients of the scores of fitting and misfitting students in test 2 with the other criteria. It also shows 95% CI for each coefficient.

Table 4.2.47 Correlations and 95% C.I. for test 2 scores with other criteria for fitting and misfitting students

	1 st term maths grade		Test 1		Final maths exam	
	Fitting	Misfitting	Fitting	Misfitting	Fitting	Misfitting
Correlation coefficient (N)	0.806 (269)	0.800 (128)	0.721 (308)	0.688 (137)	0.773 (301)	0.811 (135)
Upper (95%) limit	0.844	0.855	0.771	0.767	0.815	0.862
Lower (95% limit)	0.760	0.728	0.663	0.588	0.723	0.744

Figure 4.2.29 shows the correlation coefficients and the 95% confidence intervals in a diagrammatic form, to make comparisons easier.

Figure 4.2.29 Correlations and 95% C.I. for test 2 scores with other criteria



Again, there is absolutely no evidence of a difference in the correlations of the test scores of fitting and misfitting students with the other criteria. (Standard deviations of the scores of fitting and misfitting students were again very similar)

The students were then divided into four groups based on the results for test 1: the fitting, the misfitting by large outfit values only, the misfitting by large infit values only and the misfitting by large values from both mean square statistics. Correlations were calculated between the scores on the first test and the same criteria as above. No significant differences were found between the correlation coefficients for the four groups.

The same procedure was followed for test 2 and again no significant differences were found.

4.2.15 Comparing the internal consistency of responses for fitting and misfitting students

Cronbach’s alpha was used as a measure of the internal consistency of the raw scores in both tests. At the same time, the standard error of alpha (ASE) and 95% confidence intervals for alpha were computed using the method suggested by Iacobucci and Duhachek (2003) in order to make comparisons possible.

Test 1

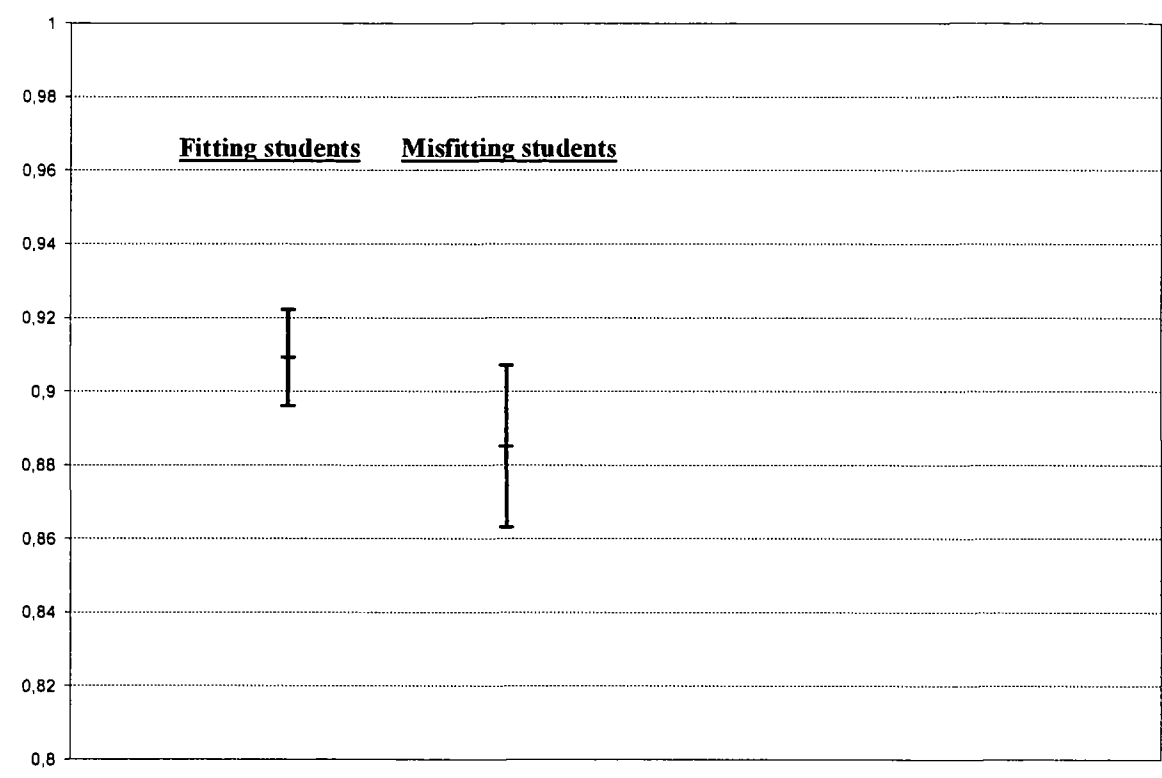
First, alpha, ASE and confidence intervals were computed for two groups, the fitting and misfitting students for the diagnostic test. Table 4.2.48 shows the results. (N is the number of students in the group, K is the number of items, ASE is the standard error of alpha and low and high are the lower and higher limits of the 95% confidence intervals for alpha)

Table 4.2.48 95% C.I. for alpha for fitting and misfitting students in the diagnostic test

Student groups	Estimate of alpha	N	K	ASE	Low	High
Fitting	0.909	413	27	0.00645	0.896	0.922
Misfitting	0.885	222	27	0.0111	0.863	0.907

Figure 4.2.30 shows the confidence intervals in a diagrammatic form. It is clear that the estimate of Cronbach’s alpha from the raw scores of the misfitting students is well below the lower limit of the 95% confidence interval of alpha from the fitting group. Therefore there is a significant difference between the alphas for the two groups of students, with the one from the misfitting students being lower.

Figure 4.2.30 95% C.I. for alpha for fitting and misfitting students in the diagnostic test



Second, alpha, ASE and confidence intervals were computed for four groups. The first was, as before, the fitting students but then the misfitting students group was divided into three groups. The one was students misfitting because of large outfit values, the second because of large infit values and the third because of a combination of large infit and outfit values. Table 4.2.49 below shows the results.

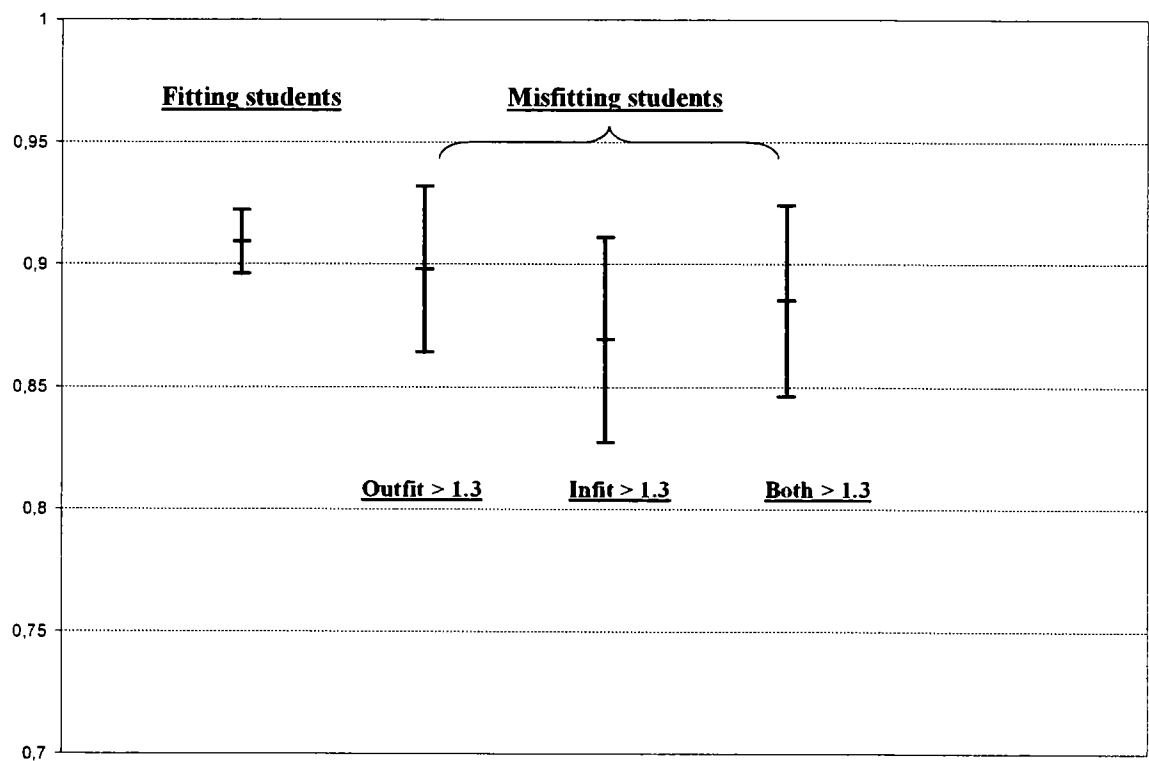
Table 4.2.49 95% C.I. for alpha for four groups of students in the diagnostic test.

Student groups	Estimate of alpha	N	K	ASE	Low	High
Fitting	0.909	413	27	0.00645	0.896	0.922
Misfitting (large outfit)	0.898	74	27	0.0171	0.864	0.932
Misfitting (large infit)	0.869	77	27	0.0215	0.827	0.911
Misfitting (large infit and outfit)	0.885	71	27	0.0197	0.846	0.924

Figure 4.2.31 shows the confidence intervals in a diagrammatic form. There are no significant differences between the alphas for the fitting students and the students who

are misfitting because of the large outfit value only (the alpha estimate for the second group is just within the 95% confidence limits of alpha for the first group). However, the alpha estimates from the misfitting students where infit is involved (either large infit only, or large infit an outfit) are well below the lower limit of the 95% confidence interval of alpha from the fitting students' group.

Figure 4.2.31 95% C.I. for alpha for four groups of students.



Test 2

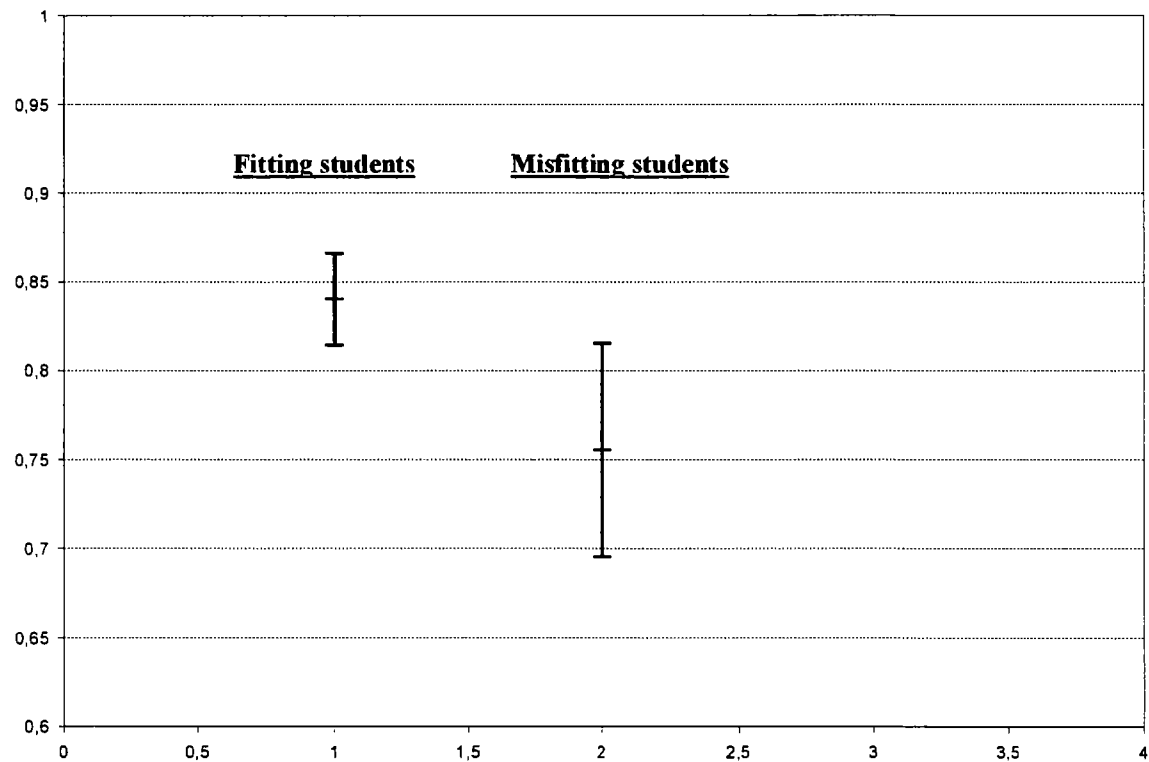
First, alpha, ASE and confidence intervals were computed for two groups, the fitting and misfitting students for test 2. Table 4.2.50 shows the results.

Table 4.2.50 95% C.I. for alpha for fitting and misfitting students in test 2

Student groups	Estimate of alpha	N	K	ASE	Low	High
Fitting	0.840	308	16	0.0133	0.814	0.866
Misfitting	0.775	137	16	0.0306	0.695	0.815

Figure 4.2.32 shows the confidence intervals in a diagrammatic form. It is clear that there are significant differences in the alpha values with the alpha estimate from the misfitting group being much lower than the one from the fitting group (the estimate from the misfitting group is well below the lower limit of the 95% confidence interval from the first group).

Figure 4.2.32 95% C.I. for alpha for fitting and misfitting students in test 2



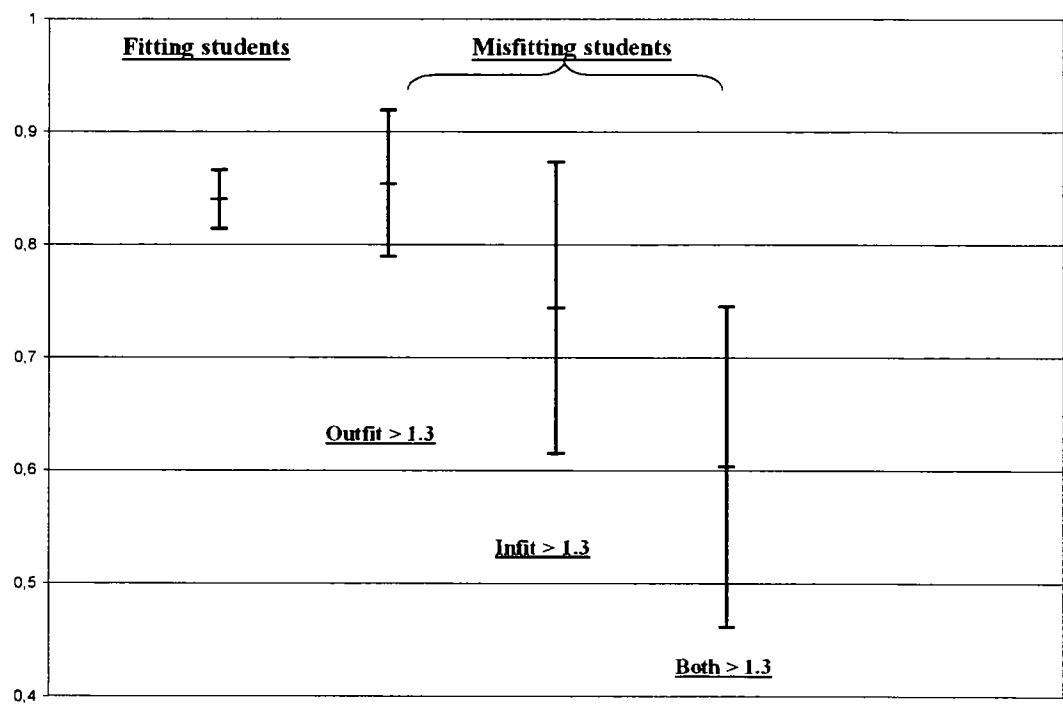
Second, alpha, ASE and confidence intervals were computed for four groups. The first was, as before, the fitting students but then the misfitting students group was again divided into three groups as before. Table 4.2.51 below shows the results.

Table 4.2.51 95% C.I. for alpha for four groups of students in test 2

Student groups	Estimate of alpha	N	K	ASE	Low	High
Fitting	0.840	308	16	0.0133	0.814	0.866
Misfitting (large outfit)	0.854	41	16	0.0333	0.789	0.919
Misfitting (large infit)	0.744	32	16	0.0661	0.615	0.873
Misfitting (large infit and outfit)	0.603	64	16	0.0725	0.461	0.745

Figure 4.2.33 shows the confidence intervals in a diagrammatic form. As before, there are no significant differences between the alphas for the fitting and misfitting-by-outfit groups. However, the alphas for the misfitting students by high infit values or high infit and outfit values are significantly lower than for the fitting students.

Figure 4.2.33 95% C.I. for alpha for four groups of students.



The same result appeared in all three maths tests used in both phases of this study, leading to the conclusion that the high infit values reduce the degree of reliability (internal consistency) of the raw scores of the students in classroom maths tests. At the same time, high outfit values do not appear to have such an effect on the reliability of the raw scores.

4.3 The Interviews

Twenty one students were interviewed in an attempt to investigate in-depth the reasons for their unexpected responses in the second maths test in phase 2. This led to a table showing the reason claimed by each student for the unexpected responses given to various items.

Then, unexpected responses were divided into two groups, the unexpected mistakes and the unexpected correct answers and explanations were given for each group, based on the students' explanations.

Finally a case of a possible misleading conclusion based on the outfit values is presented and discussed.

The sample

The sample used for the interviews consisted of 21 students from phase 2. Those were the 21 students from the researcher's school that were ranked amongst the 37 students with the most unexpected responses in the second maths test from the whole sample.

The proportion of the most misfitting students that came from the researcher's school was about 57% (21 out of 37). This proportion is very similar to the proportion of students from the whole sample that came from the researcher's school which was 59.8% (266 out of 445).

The most unexpected responses occur when the item difficulty and person ability are far apart. In such cases the outfit values tend to be very high, therefore, the students selected for the interviews, the ones with the most unexpected responses, had very large outfit values.

Six of these students were male and 15 female and they had outfit mean square values in the range 2.29 – 5.91 (5 of them had outfit > 4). Their ability estimates varied from – 0.77 to 3.31 (raw scores of 8 – 27 out of a maximum possible of 28).

4.3.1 Reasons for misfit

Table 4.3.1 below presents the reasons for the students’ unexpected responses as expressed by them during the interviews.

The first row contains the item numbers (in ascending order of difficulty). The second and third rows contain the maximum possible score and the difficulty estimate for each item respectively.

The first column contains the students’ identification numbers, the second column their ability estimates and the third their outfit values.

The remaining part of the table (rows 5 to 25 and columns 4 to 19) contains the reasons for the unexpected responses, as expressed by the students, in a coded form. The codes are as follows:

- CLS = I was Careless
- CNF = I got confused
- IGN = Didn’t know how to do it
- WGS = Wrong guessing
- NTM = No time to finish this question
- PKN-f = Prior knowledge (from the private tutor)
- PKN-t = Prior knowledge (from the class teacher)
- PKN-s = Prior knowledge (from other students)
- NEX = No explanation
- CHT = Cheating
- COR = Just got it correct
- SPRF = Special preference or knowledge
- AFCT = Possible artifact

Table 4.3.1 Reasons for the students' unexpected responses, in coded form.

Item			1	5	3	13	6	12	2	7	11	4	8	10	9	14	15	16
Max. marks			1	1	1	4	1	1	1	1	1	1	1	1	1	4	4	4
Difficulty			-2.23	-2.2	-1.43	-1.2	-0.57	-0.4	-0.12	0.07	0.61	0.62	0.69	0.75	0.86	1.05	1.69	1.82
Stud.	Abil.	Outfit																
10204	3.31	5.91				CLS												
11215	2.74	3.77				CLS						CLS						
10404	2.04	5.23	CLS															
10526	1.75	4.1		CLS					CLS									
10614	1.62	3.77		CLS			WGS											
10313	1.62	3.48	CLS															
11201	1.35	3.94				CLS		IGN										PKN-f
10419	1.35	3.31	CLS				CNF		CLS									
11214	1.21	3.3				CLS												AFCT
10722	1.05	3.05	CLS			CLS		IGN										PKN-s
11217	1.05	2.73				CLS											PKN-t	
10215	0.47	3.08				CLS									PKN-t			
11106	0.24	5.92		WGS	CNF	CNF											NEX	NEX
11223	0.24	4.63				CLS											PKN-f	
10725	0.24	3.43				CLS												PKN-s
10723	0.24	2.58			IGN													CHT
10711	0.24	2.29																PKN-f
10511	-0.27	2.78	CNF													PKN-t		
10706	-0.27	2.65	IGN															PKN-f
10211	-0.77	3.77																SPRF
10512	-0.77	2.58		CLS		NTM						NEX	CHT	PKN-f	NEX	COR		

Four of the students (11215, 10404, 10614 and 11217) claimed that test anxiety has affected their performance (but not necessarily the expectedness of their responses). A closer investigation of their test anxiety scores (Males: mean score = 20.76 s.d. = 6.87, Females: mean score = 23.88, s.d. = 7.06) showed that:

Student 11215 (Female) did not take the TAI therefore no additional confirmation of her claim can be obtained.

Student 10404 (Male) scored 33 on the TAI. His score is almost 2 standard deviations above the mean anxiety score and he was ranked approximately on the 92nd percentile.

Student 10614 (Female) scored 29 on the TAI. Her score is slightly lower than 1 standard deviation above the mean and she is ranked approximately on the 83rd percentile.

Finally, student 11217 (Female) scored 17 on the TAI. Her score is about 1 standard deviation below the mean score and she was ranked approximately on the 30th percentile.

Based on their scores, only the explanations of two students (10404 and 10614) seem to be right, that is, they were anxious for the test.

Table 4.3.1 is naturally divided into two triangular parts, the top-left and the bottom-right. These two are analysed separately, starting with the top left.

The top-left part of the table (Unexpected mistakes)

The top-left part of the table represents the unexpected mistakes made on the easier items and it is top- and left-heavy since the higher scorers and the easier items are on top and left of the table respectively. Most of these mistakes were wrong answers to easy multiple choice items, together with item 13 which was an easy and expected construct-response item, carrying a maximum possible score of 4 marks, on which some students unexpectedly lost some marks.

The main reason stated for the unexpected mistakes in these items was carelessness. Twenty out of the 30 unexpected responses were, according to the students, because of carelessness. Below there are abstracts from the students' interviews who claimed that there mistakes were simply careless.

Student 11215 (measure 2.74, raw score 26, percentile rank 96, outfit: 3.77) unexpectedly missed 1 mark (out of 4) in question 13 (measure – 1.2) and got question 4 (measure 0.62) wrong.

Interviewer: Let's see the questions now. The first question is number 4, which is asking for the roots of the equation $2x^2 + 5x = 0$. You made a mistake, you circled the wrong answer.

(She found 0 and $-5/2$ as the roots, which is correct. However, she circled 0 and -0.4 instead of 0 and -2.5). Wasn't this an easy question for you?

Student: It was, ... I basically found the correct results, I knew it was (a) but I don't know how, I circled (d),..... I probably got confused, I don't know. That was careless.

Interviewer: You mean you found the correct answer?

Student: Yes (she points in the answer sheet where she had the working out)

Interviewer: We have one more question, number 13 (she ended up with $2/6$ and instead of writing $1/3$ she wrote $1/2$)

Student: I did everything right, ... but instead of putting $1/3$ I put $1/2$ (smiles), ..., I just got confused I suppose, ... it was supposed to be $2/6$.

Interviewer: So, what you are saying is that you missed these two questions because....?

Student: CARELESS.

Another example is student 10526 (measure 1.75, raw score 21, Percentile Rank 87, Outfit: 4.10). Missed two easy multiple-choice questions, 2 (measure – 0.12) and 5 (measure – 2.2).

Interviewer: Why do you think you made a mistake here? (shows q.2)

Student: Because, ... I don't know, ...

Interviewer: Can you show me the right answer?

Student: It's this one (shows the right answer).

Interviewer: So?

Student: I saw the equation, ... and didn't realize they were not given in the right order, ... that moment I didn't realize.

.....

Interviewer: Were you in a hurry? (The researcher asked that after a long pause from the student)

Student: Yes, I wanted to finish the first ones as quickly as possible, so as to have more time on the last questions that were more difficult.

Interviewer: Ok, let's see question 5 now, can you read it please?

Student: (Reads the question).

Interviewer: What mistake did you make? (She applied the formula correctly and ended up with $9 - 8$, and instead of 1 she circled -1)

Student: Everything was correct.

Interviewer: But?

Student: But (laughs) I circled -1 instead of 1.

Interviewer: Why?

Student: Don't know why., carelessness. I have that flaw.

Ten of these 20 careless mistakes were made in question 13 (the easiest of the construct-response questions with a measure of -1.2) thus unexpectedly losing some marks.

Item 13 was:

13. Solve the equation $3x^2 + 5x - 2 = 0$.

(All this question was asking was to apply a well-known and many-times used formula to obtain the solutions of a quadratic equation).

Solution:

First step: Identify a, b and c (to use in the formula $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$)

Second step: $x = \frac{-5 \pm \sqrt{5^2 - 4(3)(-2)}}{2 \cdot 3}$

Third step: $x = \frac{-5 \pm \sqrt{5^2 - 4(3)(-2)}}{6}$

Fourth step: $x = \frac{-5 \pm \sqrt{49}}{6}$

Fifth step: $x = \frac{-5 \pm 7}{6}$

Sixth step: $x = \frac{1}{3}$ and $x = -2$.

Six out of the 10 mistakes were made from the third to the fourth step.

These students did $\sqrt{25-24}$ instead of $\sqrt{25+24}$ thus finding 1 instead of 7. It seems that whenever calculations involve the product of two negative signs, if students are not careful, they will make a mistake, no matter how able they are.

Out of the 30 unexpected mistakes in this top-left part of the table, other than the 20 identifying carelessness as the reason for their mistakes, 3 said they got confused and 4 that they didn't know what to do. Also one student claimed he did not have time to attempt question 13, which was an easy one, because he spent more time on the last ones which were the harder and that was the reason for scoring 0 marks.

Finally, 2 students blamed wrong guessing for their unexpected mistakes.

Student 10614 (measure 1.62, raw score 20, percentile rank 85, outfit: 3.77) missed two easy questions, 5 (measure - 2.2) and 6 (measure -0.57).

Interviewer: What does question 6 say?

Student: (Reads the question)

Interviewer: What was your answer?

Student: The roots are real and equal.

Interviewer: Why did you choose that?

Student: Because I didn't remember if they were equal or unequal, ...

And I just guessed they were equal. Didn't think as hard as I should.

Interviewer: Did you find the discriminant (whose value would indicate the nature of the roots)

Student: Yes

Interviewer: And you found?

Student: 20

Interviewer: So, ..., you were between (b) and (c) and you just guessed?

Student: Yes, I just figured if it is = 20 the roots must be equal.

Student 11106 (measure 0.24, raw score 12, percentile rank 48, outfit: 5.92) unexpectedly got q. 3 (measure – 1.43) and q.5 (measure – 2.2) wrong, and scored 0 marks in q. 13 (measure – 1.2). At the same time he scored 3 marks in q. 15 (measure 1.69), the second most difficult question in the test and 2 marks in q. 16 (measure 1.82), the most difficult question.

Interviewer: Question 5 now. It reads (reads the question, asking for the discriminant of the equation). You circled the answer – 17. Do you remember how you got that?

Student: Eeh, ... as I was rushing in the end (implying that he just guessed the answer)

Interviewer: You mean you just picked one at random?

Student: As I read it I put it.

Interviewer: Was it just this question you guessed the answer to?

Student: Yes. I knew that it was $\beta^2 - 4\alpha\gamma$ (He means the discriminant)

Interviewer: So, you mean you just left q. 5 last?

Student: Yes

Interviewer: (Researcher’s thought: Doubtful, it was not a difficult question).

The bottom-right part of the table (Unexpected correct answers)

The term ‘unexpected correct answer’ describes the case where a student either gets a difficult dichotomous item right or scores in a constructed response item much higher than expected.

This part of the table contains reasons, as expressed by the students, for scoring more marks than expected in some items, mainly the last 3, the more difficult construct-response items. It is obvious from the table that the last item, item 16 was the most

difficult one in the test and the one on which most of the unexpected responses were observed.

In 10 out of the 19 cases, students identified prior knowledge as the reason for the unexpectedness in their responses (this is referred to separately later), whereas two students gave no explanation as to how they scored the marks in 4 questions and 1 student claimed that she just figured it out.

Cheating

Two of the students (the ones with identification numbers 10512 and 10723) identified, or admitted rather, that cheating was the reason for their unexpected correct responses.

Student 10723 (measure 0.24, raw score 12, percentile rank 48, outfit 2.58) got question 3 (measure – 1.43) wrong and at the same time she scored almost full marks (3 out of 4) in q. 16 (measure 1.82), the most difficult question in the test.

Interviewer: This question now (shows q. 16). It was the most difficult of the test. You have almost solved it completely.

Student: (Smiles)

Interviewer: That smile means something, doesn't it? Tell me.

Student: ok, ...

Interviewer: I know what you are going to say (having prior knowledge in mind), so please say it, don't worry, whatever you say, as I explained before, is between us.

Student: What should I say? (Smiles)

Interviewer: Tell me, sincerely, how you got this question right.

Student: Ok, I saw this question.

Interviewer: You mean from somebody else during the test?

Student: Yes, from the person in front, ... up to here , and then I continued by myself with D, we all know that. (She means she went on to solve the quadratic equation by herself, and made a small mistake on the calculations with a minus sign).

Student 10512 (measure - 0.77, raw score 8, percentile rank 22, outfit 2.58) scored 0 marks in q. 5 (measure - 2.2) and 0 out of 4 in q. 13 (measure - 1.2). At the same time she scored 1 mark (that is she got the right answer) in q. 4, 8, 9 and 10 (measures 0.62, 0.69, 0.75 and 0.86 respectively) all well above her ability estimate. Also, she scored 2 out of the 4 possible marks in q. 14 (measure 1.05) the third most difficult question in the test.

Interviewer: Ok, let's see question 8 now, it was one of the more difficult questions; many students got up to $4k = 9$ and then selected the wrong value for k .

Student: I wasn't quite sure about that and got a little help from the girl behind me. (Laughs)

Interviewer: So, the girl behind you told you the answer?

Student: No, she saw that I was doing it wrong and told me how to do it.

Interviewer: Oh, I see.

Student: Just there she helped me. (Saying, before being asked, that she got help in that question only)

Special preference or special knowledge

Another possible reason for the unexpected correct answers was special preference, or special knowledge, in a certain topic. One of the weaker students (student 10211, measure - 0.77, percentile rank 22, outfit 3.77), managed to start question 16 right and scored 2 out of the 4 marks. Her explanation about this was special preference or special knowledge.

Below there is an extract from her interview

Interviewer: Ok, up to here (shows where she stopped in her answer to q.16, she managed to get the first part right) you did it correctly? Did you understand the question?

Student: Not quite, until I had the time to think about it, the bell rang.

Interviewer: You mean you could have finished it if you had enough time?

Student: If I had more time, yes.

Interviewer: Even though it was the most difficult question? Most students didn't manage to do it.

Student: No, because these questions with length, width, ... I like them much more than roots.

Interviewer: You mean because there was a little geometry in the question?

Student: Geometry is my God in maths, that's why I could do this question.

Interviewer: Ok, thank you very much.

Her explanation for this unexpected answer was simply that she likes (or is more able or both) geometry much more than algebra and question 16 was the most original item, combining a little geometry with the algebra. She managed to get the first part of the question right, the one which was based on knowledge of simple geometry together with the algebraic calculations required, but then she could not finish the second part, which required forming and solving a quadratic equation.

Prior Knowledge

Prior knowledge was the most frequent explanation for the unexpected correct answers. In 10 out of the 19 cases encountered in the sample the reason behind unexpected correct answers was prior knowledge. However, from the students' answers during the interviews it became obvious that 3 different types of prior knowledge could be identified.

(i) *Prior knowledge from the private tutors*

Five of these 10 students (the ones with identification numbers 11201, 11223, 10711, 10511 and 10512) attributed their prior knowledge to specific questions to their private tutors who either had the insight or the information to give their tutees similar items with the ones in the test for practice. Below there are extracts from the interviews of two of such students.

Student 11201 (measure 1.35, raw score 18, percentile rank 78, outfit 3.94) unexpectedly lost 2 marks in question 13 (measure – 1.2) and got question 12 wrong

(measure – 0.4). At the same time he got full marks on the most difficult question, 16 (measure 1.82).

Interviewer: Question 16 now. In this one you scored full marks. This question was considered the most difficult. How did you figure it out?

Student: To be honest, ..., we asked students from other classes, ... they told us that there would be a question like that, ... I took it to the private lessons, and we solved it there, more or less, that's how I understood it. It wasn't exactly the same one, but I did it exactly the way it was explained to me, it wasn't difficult.

Interviewer: Ok, thank you very much and especially for your sincerity.

Student 10711 (measure 0.24, raw score 12, percentile rank 48, outfit 2.29) scored almost full marks (3 out of 4) in q. 16 (measure 1.82), the most difficult question in the test.

She doesn't really like maths, but since she started private tuition she does better. She was nervous before the test but that did not affect her performance.

Interviewer: Let's see q.16. It was the most difficult question in the test and yet you managed to score 3 out of the 4 marks. At the same time you lost marks on much easier questions. Why do you think?

Student: Because we did it at the institute (She means the private lessons).

Interviewer: The exact same one?

Student: No, similar.

Interviewer: Did any of your classmates have the test, and took it to the institute?

Student: No, the question was similar to 16.

(ii) *Prior knowledge from other students*

Students 10722 and 10725 attributed prior knowledge to information they received from other students from other classes who had taken the test earlier and passed onto them

some questions. Given the originality of the last and most difficult question and the fact that it was the only one containing a diagram it was easier to remember and to pass on.

Student 10722 (measure 1.05, raw score 16, percentile rank 70, outfit: 3.05) unexpectedly scored 0 marks in questions 1 (measure - 2.23), 12 (measure - 0.4) and lost 1 mark in question 13 (measure - 1.2). At the same time she managed to score 3 (out of a possible 4) marks in question 16 (measure 1.82), the hardest question on the test.

Interviewer: Well, let's go on to q.16, this was a difficult question and yet you got it almost right. How come?

Student: I put 6 instead of 16 (in the diagram it was 16, but she wrote 6, that looks like a careless mistake)

Interviewer: How did you figure that question out? Tell me, honestly, if it was so difficult, how did you get it almost right?

Student: Ok, ..., I was warned by other students who took the test before us, about a question with areas, and so, ... like, ... They explained the general idea of the question to me, how to start it, and then I tried to finish it myself.

Interviewer: So, some students knew what the last question was?

Student: Yes.

Interviewer: Thank you very much, especially for your sincerity.

Student 10725 (measure 0.24, raw score 12, percentile rank 48, outfit: 3.43) lost 3 marks in q. 13 (measure - 1.2) and at the same time she scored almost full marks (3 out of 4) in q. 16 (measure 1.82), the most difficult question in the test.

Interviewer: Let's see this one now (shows q. 16). It was a very difficult question, very few students got it right, and you solved it. Ok, if we ignore a minor mistake in the end you solved it. How did you figure it out? I mean these two (q. 14 and 15) were easier and you didn't get them. Tell me, sincerely.

Student: I don't know, it looked easier to me, because it involves areas that we did previously, ... it looked easier.

Interviewer: Did you by any chance hear anything about this question (before the test)?

Student: Some students brought it just before the test, don't know, from their institute I suppose (she means private lessons).

Interviewer: The exact same question?

Student: No, no, like a similar one, like it could be in (the test)

Interviewer: So, basically you saw it and you realized what you have to do with the areas?

Student: No, they told me you should do so and so and then you are on your own.

Interviewer: I see. So some students had an idea of what the question would be and they said how it should be started and then everyone could try from then on to finish it?

Student: Yes.

Interviewer: Ok, thank you very much.

(iii) Prior knowledge from the class teacher

Students 10511, 10215 and 10217 attributed prior knowledge to their class teachers who, in their attempt to prepare their students better for the test, gave hints for questions 14 and 15.

Student 11217 (measure 1.05, raw score 16, percentile rank 70, outfit 2.73) lost 2 marks (out of 4) in q. 13 (measure – 1.2). At the same time she managed to score full marks in q. 15 (measure 1.69) the second most difficult question in the test.

Interviewer: Let's see q. 15 now. How did you figure it out? It was a difficult question.

Student: Eeeeh, Mr. (the maths teacher) joined together lots of questions, so,.... I don't know how I got it. I thought, ..., I don't know how I thought, ... and got it.

Interviewer: He joined them together? What do you mean? You did a similar question?

Student: No, similar parts. He divided the question into smaller parts, eeeeh, ... it was like different separated questions. Let's say we did one time this part, another time the other part, .. and so on.

Interviewer: You mean you did this question, or similar to this, piece by piece?

Student: Yes.

Student 10215 (measure 0.47, raw score 13, percentile rank 54, outfit 3.08) unexpectedly lost 3 mark in q. 13 (measure - 1.2) and scored full marks in q. 14 (measure 1.05) which was the 3rd most difficult question in the test.

Interviewer: This question now (shows q. 14). How did you get this one right?

Student: Opposite numbers means their sum is 0. Therefore $S = 0$, and then I solved it.

Interviewer: So, you didn't find it difficult?

Student: Not at all. We did questions like this in the class so It wasn't difficult.

Carelessness seems to be the most important reason for unexpected mistakes in the maths tests. At the same time prior knowledge seems to be the most important reason for the unexpected correct answers. Possible explanations for this include the following: This test was a classroom maths test. It was not taken by all classes at the same time. Teachers taught their classes in their own pace, therefore they administered the test when they had finished the chapter on quadratic equations. The tests were administered within a period of about one week separately in each school.

Teachers teaching more than one class however, managed to administer the tests in the same day.

Given also the fact that the students in one school live in the same area of the town, and that some of them attend private lessons at the same tutors, it was not too difficult for their tutors to figure out, after a couple of days what some possible test questions would be like. Especially after a couple of private sessions, where most of the students would be describing the last and difficult (and easy to remember because it contained a diagram) question to them.

4.3.2 Possible artifact (inflated outfit)

In some cases one or two careless mistakes lower the raw score, and consequently the ability is underestimated thus making solutions to more difficult questions seem more unexpected too. (Therefore, the outfit could be inflated by the wrong identification of some correct responses to difficult questions as unexpected, i.e. higher residuals).

For example, student 11214 lost very carelessly 2 marks in question 13 (measure – 1.2), thus getting an ability estimate of 1.21 (percentile rank = 75). Had she answered question 13 correctly her ability estimate would have gone to 1.49 (percentile rank = 82), thus making the correct solution to the last question (measure 1.82) not so unexpected.

This point is explained further with an example and the help of two tables. Table 4.3.2 shows the responses of student 11214 and table 4.3.3 the responses of a hypothetical student 91214 whose responses are the same as those of student 11214 except from item 13 where the latter avoids the careless mistake and scores 4 marks instead of 2. All the figures in the tables are rounded off to 2 decimal places.

The first row in the two tables contains the item numbers (in entry order) and the second row the item measures.

The third and fourth rows contain the observed and the expected (based on the Rasch model) scores of the two students.

The fifth row contains the residuals, which are obtained by subtracting the expected from the observed score. The model variance around the expected value is shown in the sixth row and the standardized residuals (residuals divided by the equivalent standard deviations) are shown in the seventh row.

The sum of the squared standardized residuals divided by 16 (the number of items) gives the value of the outfit mean square. That is, outfit is the mean of the squared standardized residuals.

Finally, the last row of the two tables shows the contribution of the response to each item to the outfit value. The sum of all the numbers in the last row gives again the value of the outfit.

Table 4.3.2 Scores and residuals for students 11214 (with the observed score of 2 marks in item 13)

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Item measure	-2.23	-0.12	-1.43	0.62	-2.20	-0.57	0.07	0.69	0.86	0.75	0.61	-0.40	-1.20	1.05	1.69	1.82
Observed score	1	1	1	0	1	1	1	0	0	0	0	1	2	1	3	4
Expected Score	0.97	0.79	0.93	0.64	0.97	0.86	0.76	0.63	0.59	0.61	0.65	0.83	3.90	1.98	0.97	0.91
Residual (Obs – Exp)	0.03	0.21	0.07	-0.64	0.03	0.14	0.24	-0.63	-0.59	-0.61	-0.65	0.17	-1.90	-0.98	2.03	3.09
Model variance	0.03	0.17	0.06	0.23	0.03	0.12	0.18	0.23	0.24	0.24	0.23	0.14	0.11	1.88	1.48	1.37
Standardised residual	0.18	0.51	0.27	-1.34	0.18	0.41	0.57	-1.30	-1.19	-1.26	-1.35	0.45	-5.76	-0.72	1.67	2.64
Contribution to outfit	0.00	0.02	0.00	0.11	0.00	0.01	0.02	0.11	0.09	0.10	0.11	0.01	2.07	0.03	0.17	0.43

Outfit = 3.30

Table 4.3.3 Scores and residuals for students 91214 (with a hypothetical score of 4 marks in item 13)

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Item measure	-2.23	-0.12	-1.43	0.62	-2.20	-0.57	0.07	0.69	0.86	0.75	0.61	-0.40	-1.20	1.05	1.69	1.82
Observed score	1	1	1	0	1	1	1	0	0	0	0	1	4	1	3	4
Expected Score	0.98	0.83	0.95	0.70	0.98	0.89	0.81	0.69	0.65	0.68	0.71	0.87	3.93	2.52	1.46	1.36
Residual (Obs – Exp)	0.02	0.17	0.05	-0.70	0.02	0.11	0.19	-0.69	-0.65	-0.68	-0.71	0.13	0.07	-1.52	1.54	2.64
Model variance	0.02	0.14	0.05	0.21	0.02	0.10	0.16	0.21	0.23	0.22	0.21	0.11	0.08	1.89	2.05	1.86
Standardised residual	0.16	0.45	0.23	-1.54	0.16	0.36	0.49	-1.49	-1.37	-1.45	-1.55	0.39	0.25	-1.11	1.07	1.93
Contribution to outfit	0.00	0.01	0.00	0.15	0.00	0.01	0.02	0.14	0.12	0.13	0.15	0.01	0.00	0.08	0.07	0.23

Outfit = 1.12

Winsteps (Linacre, 2005) places the response to any item with standardised residual greater than 2 or smaller than -2 in the 'most unexpected' category.

Student 11214 has a total score of 17 out of 28, an ability estimate of 1.21 (75th percentile) and an outfit of 3.30. Her responses to items 13 (standardised residual - 5.76) and 16 (standardised residual 2.64) are flagged as unexpected. These are the two highlighted columns in the first tables.

In item 13 student 11214 scored 2 marks. Based on her ability, the expected score on this item was 3.90 thus yielding a rather large residual of -1.90 and a standardised residual of -5.76 . This standardised residual contributes 2.07 to the outfit value, making her response pattern for the whole test unexpected and suspect. (For the purposes of this study a cut-off score of 1.3 was used for both the outfit and infit values).

At the same time, student 11214 scored 4 marks in item 16 where her expected score was only 0.91. This has yielded a residual of 3.09 and a standardised residual of 2.64 thus flagging the response to this item as one of the 'most unexpected'. Furthermore, the contribution of this standardised residual to the overall outfit is 0.43, which is 33% of the outfit cut-off value of 1.3.

The expected score in item 16 is based on the ability of the student which has already been estimated with the total score of 17 out of 28. However, this student has very carelessly lost two marks thus yielding a lower than the true estimate of her ability. (She is, as expressed by herself during the interview, a very capable student, who likes maths very much and feels very confident in the subject. She does not make careless mistakes too often).

Student 91214 (the hypothetical student) has the same response pattern as student 11214, in 15 out of the 16 items, but in item 13 she didn't make the careless mistake and scored 4 marks. This has the following consequences:

- The total score is now 19 out of 28.
- The ability estimate is now 1.49 (82nd percentile).

- The standardised residual on item 13 is now only 0.25 (since the expected score of 3.93 is almost the same as the observed score of 4) contributing 0.00 to the outfit value.
- More importantly however, the residual in item 16 is now 2.64 (and not 3.09), the standardised residual is 1.93 (thus the response is not flagged as most unexpected) and the contribution to the outfit value is only 0.23 (a decrease of 47%) which now is less than 18% of the outfit cut-off value of 1.3.
- Finally the outfit value has decreased from 3.30 to 1.12 and within the acceptable range.

Comparing the responses of the two students one can see that if the careless mistake in item 13 had not occurred, the response to item 16 would have no longer been identified as one in the 'most unexpected' category.

Therefore it seems that if a student loses a couple of marks because of careless mistakes, that would not only result in a lower ability estimate (than the true ability) but also in a higher chance of misidentifying responses to difficult items as unexpected and inflating the value of the outfit.

The large difference in the standardised residuals also has an impact on the infit. However this impact is not as high as on the outfit. In the example mentioned above, the infit of student 11214 was 3.02 whereas the corresponding value for the hypothetical student (91214) was 1.86. There is a large difference in the two values; however both indicate an unexpected response string, because of the weight placed on the responses to the on-target items.

With regard to the infit and outfit values of item 13 or 16, given the large number of students involved, there is no real difference in their values.

4.4 Investigating outfit and infit

This investigation was undertaken in an attempt to explain why the internal consistencies of misfitting students with high infit were of a lower degree than those of fitting students or misfitting students with high outfit.

4.4.1 The effect of test length on the outfit of a response string with one unexpected answer

This chapter presents an investigation into the effect of test length on the outfit of a response string with one unexpected answer. In particular the researcher investigated which test lengths (up to how many items) would cause the outfit mean square to exceed the cut-off score (1.2 or 1.3) if the response string contains only one unexpected response.

Unexpected answer on an item 3 logits away from the student's ability

For this investigation a theoretical set of a varying number of dichotomous items was used with their difficulties uniformly spread in the range from – 2.0 to 2.0 logits.

First, 9 items were evenly spread (having a mean difficulty of 0) in the range and a hypothetical student of ability 1 was used. This student had a deterministic response string, which in psychometrics is called the Guttman response string (correct answers for all items with difficulties up to and including 1 logit, and wrong answers for items with difficulties above 1 logit). Table 4.4.1 below shows the deterministic response string for this first case with the 9 items.

Table 4.4.1 Items and responses

Items	1	2	3	4	5	6	7	8	9
Item difficulties	-2.0	-1.5	-1.0	-0.5	0	0.5	1.0	1.5	2.0
Responses	1	1	1	1	1	1	1	0	0

Outfit and infit mean squares were calculated. Then the response on item 1 (3 logits away from the ability estimate of 1 logit) was changed into 0 making it an unexpected response (for example, a careless mistake).

Outfit and infit mean squares were calculated again for this response string with the one unexpected answer.

Having finished that, the procedure was repeated 5 more times with varying test lengths (11, 17, 21, 25 and 27 items) spread in the same range of item difficulties from -2.0 to 2.0 logits.

Table 4.4.2 below shows the outfit and infit mean squares for the deterministic response string (D – response string) and for the one with 1 unexpected answer (U – response string) for the different test lengths.

Table 4.4.2 Outfit and infit for persons for the different test lengths

Test length (N of items)	D – response string		U – response string	
	Outfit	Infit	Outfit	Infit
9	0.382	0.496	2.608	1.106
11	0.374	0.481	2.196	0.976
17	0.387	0.494	1.565	0.811
21	0.388	0.494	1.342	0.750
25	0.401	0.501	1.202	0.712
27	0.389	0.497	1.131	0.696

The two columns under the D – response string contain the calculated fit statistics for the deterministic response strings, and as expected, they show overfit to the Rasch model.

The last column shows the effect of test length on infit, and as the test length increases, students with response strings with only one off-targeted unexpected answer tend to become overfitting. Even in a short test containing as few as 9 items, one unexpected off-targeted answer only slightly affected the infit. (With shorter tests infit would go over 1.2 or 1.3 as well).

The column in bold, under U – response string, shows the effect of test length on the outfit. It is obvious that the shorter the test the higher the impact of the unexpected off-targeted answer on the outfit value.

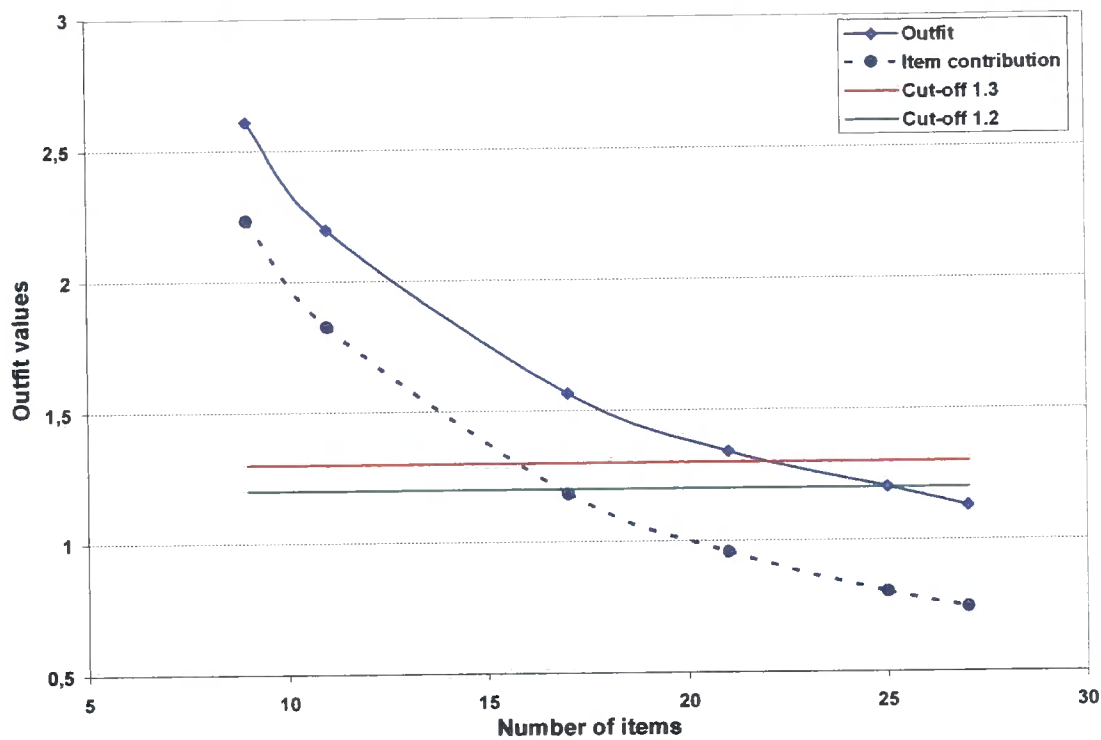
Figure 4.4.1 shows how the outfit values vary as test length increases. It also shows the contribution of the item with the unexpected answer to the overall outfit value. Item contribution to the outfit value is simply the impact of the specific unexpected response to the overall outfit value. Outfit is the average of the squared standardised residuals. Therefore, if one divides the squared standardised residual of each response by N (number of test items) one can find how much each response contributes to the overall outfit value.

Finally, the figure contains the cut-off lines at values of 1.2 and 1.3.

One can see that if a test has a length of 22 items or less the effect of the one unexpected answer (at an item 3 logits away from the student's ability) is that outfit exceeds the 1.3 cut-off value, categorizing this response as aberrant. If 1.2 is used as the cut-off value for the outfit, then the equivalent test length is 25 items or less.

The contribution to the outfit value of the one unexpected answer is large: from 86% on the 9-item test to 66% on the 27-item test. The implication of this contribution is that for a test of length 15 or less, in the 1.3 cut-off value case (16 or less when 1.2 is used), the square standardised residual of the specific item by itself makes the whole response string misfitting, (i.e. even if the squared standardised residuals of all other responses on the other items were zeros, the outfit value would still exceed the cut-off value).

Figure 4.4.1 Effect of test length on outfit



A further investigation on the squared standardised residuals was conducted showing that: if $|B_n - D_i| = \text{constant}$ for any values of B_n and D_i , where B_n and D_i are the student's ability and item's difficulty, the squared standardised residuals will be the same and therefore the contribution to the outfit the same, for a specific test length.

For example, table 4.4.3 below shows the calculation of the squared standardised residuals for 4 items with unexpected responses from 4 students (one for each item) and in all the cases the difference between ability and difficulty is ± 3 .

The probability of success (p) is calculated using the Rasch model formula. The expected score $E(X)$ is calculated using:

$$E(X) = \sum_{\text{all } x} x \cdot P(X = x), \text{ where } x = 0, 1 \Rightarrow E(X) = 0 \cdot (1 - p) + 1 \cdot p = p \text{ (Proof 1)}$$

Table 4.4.3 Calculation of the squared standardised residual

	Unexpected Responses			
	Wrong		Right	
Student ability (B_n)	1	2	- 1	0
Item difficulty (D_i)	- 2	- 1	2	3
$B_n - D_i$	3	3	- 3	- 3
Observed score	0	0	1	1
Expected score (= p, probability of success)	0.953	0.953	0.047	0.047
Variance = $p(1-p)$	0.045	0.045	0.045	0.045
Standard deviation	0.213	0.213	0.213	0.213
Residual (Observed – Expected)	- 0.953	- 0.953	0.953	0.953
Standardised Residual = $\frac{\text{Residual}}{\text{Standard deviation}}$	- 4.482	- 4.482	4.482	4.482
Squared Standardised Residual	20.086	20.086	20.086	20.086
Contribution to outfit = $\frac{\text{Squared st. residual}}{\text{Testlength } N}$	$\frac{20.086}{N}$	$\frac{20.086}{N}$	$\frac{20.086}{N}$	$\frac{20.086}{N}$

It is obvious then that the curve representing the contribution to the outfit (C), in figure 4.4.1 above, is in fact the graph of the function $C = \frac{20.086}{N}$, for various values of N.

More generally, it is the graph of the function:

$$C = \frac{\text{Squared st. residual}}{\text{Testlength } N}$$

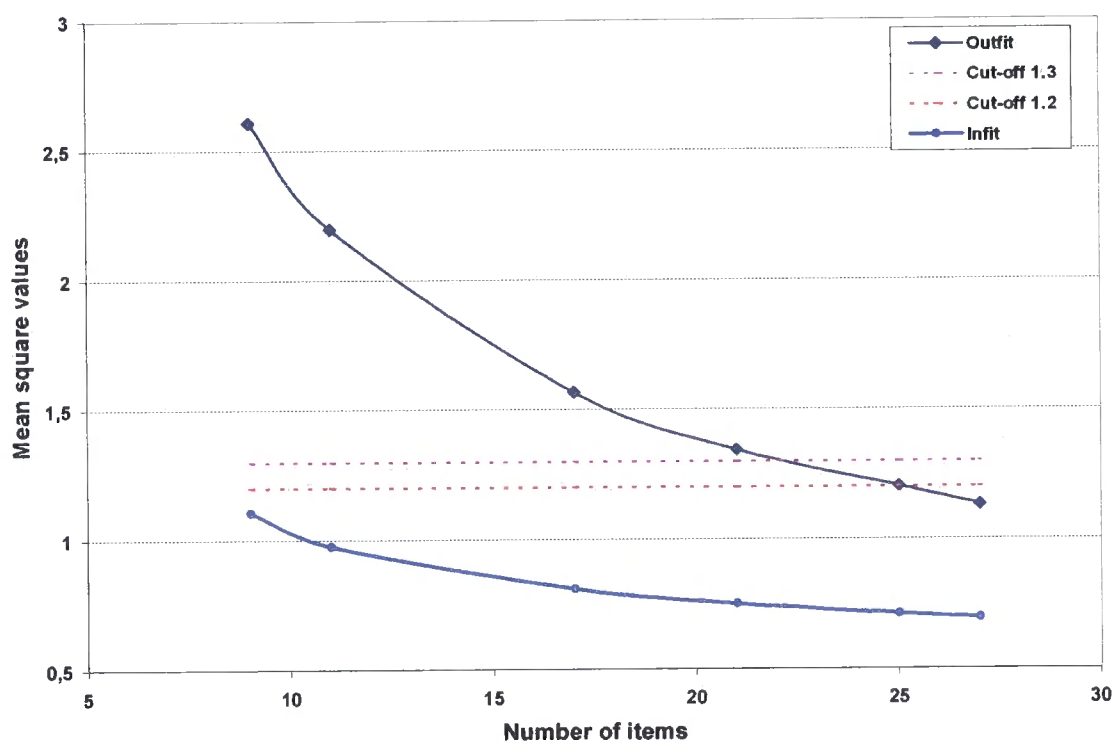
If one wants to find the number of items below which the contribution of an unexpected response is greater than the cut-off value of 1.3 or 1.2, all one has to do is to solve one of the inequalities:

$$\frac{\text{Squared st. residual}}{\text{Testlength } N} > 1.3 \quad \text{and} \quad \frac{\text{Squared st. residual}}{\text{Testlength } N} > 1.2.$$

For example, in the above case where the item is 3 logits away from the student's ability, solving the inequality $\frac{20.086}{N} > 1.3$, gives $N < 15.45$. This can be interpreted as: If there is an unexpected response (right or wrong) on an item with difficulty 3 logits away from the student's ability, then this response by itself will make the outfit value exceed the cut-off value of 1.3 (thus categorizing the whole response string as aberrant) for any test with 15 items or less.

Figure 4.4.2 is similar to figure 4.4.1 but includes the infit as well. It is obvious that the infit is only slightly affected by one off-targeted unexpected response for reasonable test lengths. It seems that infit would exceed the 1.3 or 1.2 cut-off values in very short tests containing less than 8 items.

Figure 4.4.2 Effect of test length on outfit and infit



Unexpected answer on an item 4 logits away from the student's ability

For this investigation a theoretical set of dichotomous items was used with their difficulties following a rectangular distribution in the range from -2.5 to 2.5 logits. The range was slightly larger than before in order to accommodate for the difference of 4 logits between ability and difficulty.

The same procedure was followed, as in the previous example, but since such a response is much more unexpected than before (giving much larger squared standardised residuals) larger test lengths were needed.

First, 11 items were evenly spread (having a mean difficulty of 0) in the range and a hypothetical student of ability 2 was used. Table 4.4.3 below shows the deterministic response string for this first case with the 11 items.

Table 4.4.4 Items and responses

Items	1	2	3	4	5	6	7	8	9	10	11
Item difficulties	-2.5	-2.0	-1.5	-1.0	-0.5	0	0.5	1.0	1.5	2.0	2.5
Responses	1	1	1	1	1	1	1	1	1	1	0

Outfit and infit mean squares were calculated.
Then the response on item 2 (4 logits away from the ability estimate) was changed into 0 making it an unexpected response.
Outfit and infit mean squares were calculated again for this response string with the one unexpected answer.

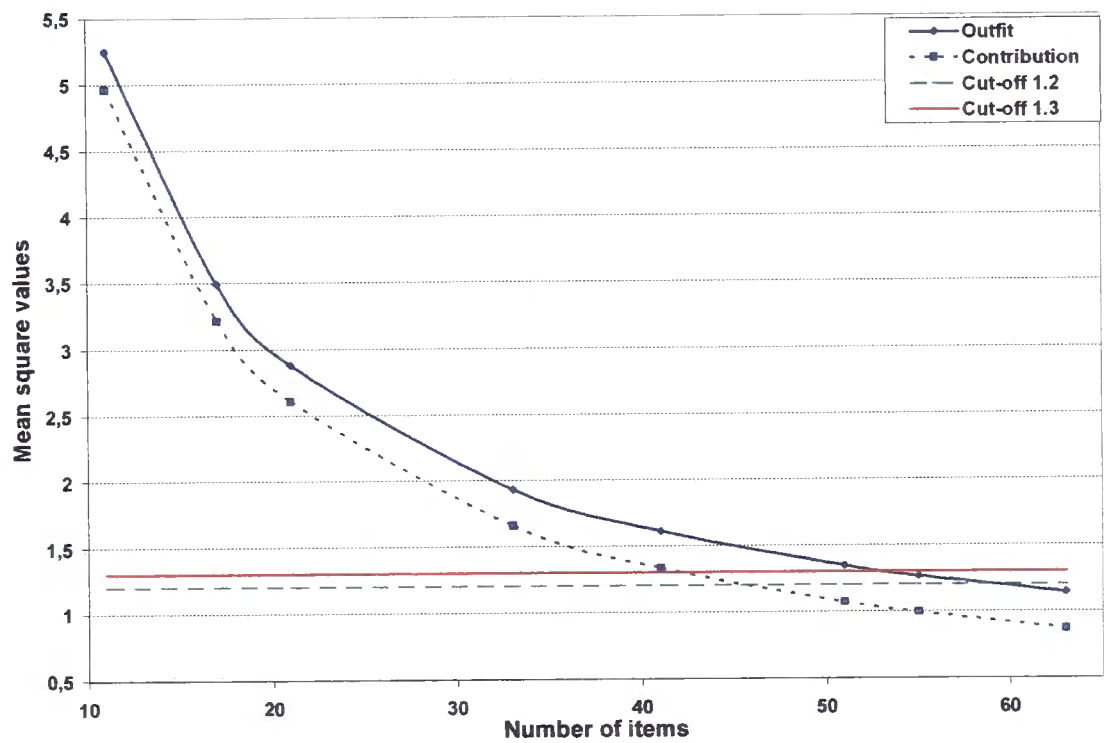
The procedure was repeated 7 more times with varying test lengths (17, 21, 33, 41, 51, 55 and 63 items) spread in the same range of item difficulties from -2.5 to 2.5 logits. Table 4.4.5 below shows the outfit and infit mean squares for the response strings with the 1 unexpected answer (U – response string) for the different test lengths (The infit and outfit for the deterministic response strings were also calculated and were for the outfit from 0.285 down to 0.276 and for the infit 0.495 down to 0.472, both overfitting the model and decreasing as N increased)

Table 4.4.5 Outfit and infit for the different test lengths

Test length (N of items)	U – response string	
	Outfit	Infit
11	5.246	1.212
17	3.492	0.950
21	2.878	0.857
33	1.931	0.714
41	1.612	0.675
51	1.348	0.631
55	1.267	0.613
63	1.142	0.597

Figure 4.4.3 shows again how the outfit values vary as test length increases. It also shows the contribution of the item with the unexpected answer to the overall outfit value.

Figure 4.4.3 Effect of test length on outfit



One can see that if a test has a length of about 53 items or less the effect of the one unexpected answer (at an item 4 logits away from the student's ability) is that outfit exceeds the 1.3 cut-off value, categorizing this response as aberrant. If 1.2 is used as the cut-off value for the outfit, then the equivalent test length is about 60 items or less.

The squared standardised residual of the unexpected response was 54.598. The contribution to the outfit (C) graph is simply (as explained in the previous example) the graph of the function:

$$C = \frac{54.598}{N}.$$

Solving the inequality $\frac{54.598}{N} > 1.3$ gives $N = 42$, and (1.3, 42) is the point on the figure where the cut-off line at 1.3 meets the contribution to outfit graph. When 1.2 is used as the cut-off value then $N = 45.5$.

This can be interpreted again as: If there is an unexpected response (right or wrong) on an item with difficulty 4 logits away from the student's ability, then this response by itself will make the outfit value exceed the cut-off value of 1.3 (thus categorizing the whole response string as aberrant) for any test with about 42 items or less.

Therefore, in the case of this study, where classroom tests are involved with smaller numbers of items, 1 unexpected response on an item 4 logits away from the student's ability would definitely categorise the student's response string as aberrant and the student as misfitting.

A good real example, to illustrate the effect of just one unexpected response, is student 10404 in test 2 of phase 2 (this student is included in the group that were interviewed in phase 2).

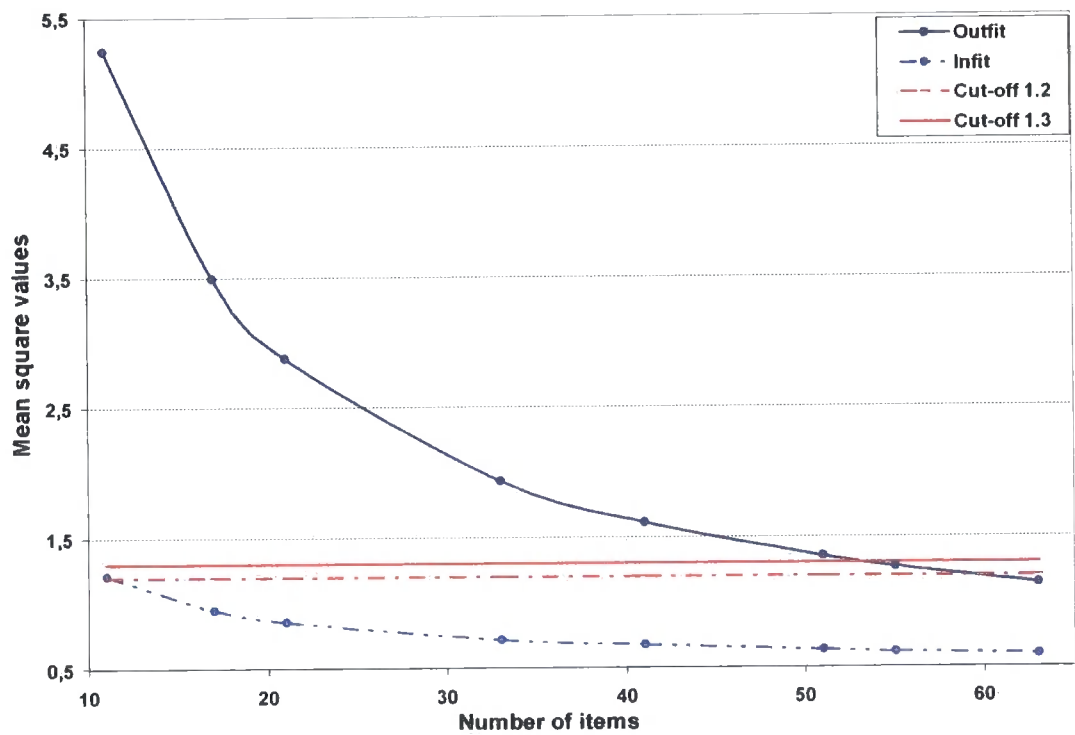
This student took a 16-item test, and his ability was estimated at 2.04. However, he missed a very easy dichotomous item, item 1, with a measure of -2.23 (4.27 logits away from the student's ability) thus having an **outfit of 5.23**.

The squared standardised residuals for this unexpected answer (the probability of success on that item was 0.986 and the probability of failure only 0.014) were 71.522.

The contribution of that answer to the outfit was 4.470 making his response string extremely aberrant.

Figure 4.4.4 is similar to figure 4.4.3 but includes the infit as well. It is obvious that the infit is not affected in a significant way by one off-targeted unexpected response for reasonable test lengths. It seems that infit would exceed the 1.3 or 1.2 cut-off values in very short tests containing less than 10 or 11 items.

Figure 4.4.4 Effect of test length on outfit and infit



4.4.2 A formula for the contribution of an unexpected response to the outfit

The contribution of any unexpected response to the outfit is given by:

$$C = \frac{\text{Squared st. residual}}{\text{Testlength } N}$$

Residual = $O_i - E_i$ where O_i and E_i are the observed and expected responses to item i respectively.

The Standardised Residual is the residual divided by the model standard deviation and

$$\text{Squared Standardised Residual} = \frac{(O_i - E_i)^2}{\text{Variance}} = \frac{(O_i - p)^2}{p(1-p)} \text{ since } E_i = p \text{ (see proof 1)}$$

Therefore if $B_n - D_i = \Delta_{ni}$ or for simplifying the working out we let $\Delta_{ni} = \Delta$

$$C = \frac{(O_i - p)^2}{Np(1-p)} = \frac{\left(O_i - \frac{e^\Delta}{1+e^\Delta}\right)^2}{N \cdot \frac{e^\Delta}{1+e^\Delta} \cdot \left(1 - \frac{e^\Delta}{1+e^\Delta}\right)}$$

For an unexpected wrong answer $O_i = 0$ and $B_n > D_i \Rightarrow B_n - D_i > 0 \Rightarrow \Delta > 0$, therefore

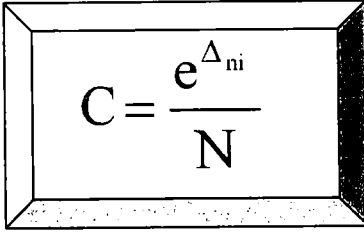
$$C = \frac{\left(0 - \frac{e^\Delta}{1+e^\Delta}\right)^2}{N \cdot \frac{e^\Delta}{1+e^\Delta} \cdot \left(1 - \frac{e^\Delta}{1+e^\Delta}\right)} = \frac{\left(\frac{e^\Delta}{1+e^\Delta}\right)^2}{N \cdot \frac{e^\Delta}{1+e^\Delta} \cdot \left(\frac{1+e^\Delta - e^\Delta}{1+e^\Delta}\right)} = \frac{\left(\frac{e^\Delta}{1+e^\Delta}\right)^2}{N \cdot \frac{e^\Delta}{(1+e^\Delta)^2}} = \frac{e^\Delta}{N}$$

For an unexpected right answer $O_i = 1$ and $B_n < D_i \Rightarrow B_n - D_i < 0 \Rightarrow \Delta < 0$, therefore

$$C = \frac{\left(1 - \frac{e^\Delta}{1 + e^\Delta}\right)^2}{N \cdot \frac{e^\Delta}{1 + e^\Delta} \cdot \left(1 - \frac{e^\Delta}{1 + e^\Delta}\right)} = \frac{\left(1 - \frac{e^\Delta}{1 + e^\Delta}\right)}{N \cdot \frac{e^\Delta}{1 + e^\Delta}} = \frac{\left(\frac{1 + e^\Delta - e^\Delta}{1 + e^\Delta}\right)}{N \cdot \frac{e^\Delta}{1 + e^\Delta}} = \frac{1}{N \cdot e^\Delta}$$

However since $\frac{1}{e^{-x}} = e^x$ the formula for an unexpected wrong response is exactly the same as the formula for an unexpected correct response and

if $|B_n - D_i| = \Delta_{ni}$, or, in words, if Δ_{ni} is the positive difference between the ability of person n and the difficulty of item i then the contribution of an unexpected response to the outfit mean square is given by the formula:



$$C = \frac{e^{\Delta_{ni}}}{N}$$

(Formula 1)

Using this formula one can investigate the minimum absolute difference between person ability and item difficulty $(|B_n - D_i| = \Delta_{ni})$ needed to make the contribution of a single unexpected response to the outfit exceed any cut-off value.

This means the minimum value of Δ_{ni} for which just one unexpected response in an N -item test would make the response string aberrant.

Rearranging the formula gives

$$e^{\Delta_{ni}} = N \cdot C \Rightarrow \Delta_{ni} = \ln(N \cdot C)$$

(Formula 2)

For example, in the case of a 20-item test the contribution of an unexpected response to the outfit will be greater than 1.3 for:

$$\Delta_{ni} = \ln(N \cdot C) = \ln(20 \cdot 1.3) = \ln(26) = 3.258$$

Therefore, an unexpected response to an item with difficulty 3.258 logits away from the person’s ability will contribute to the outfit value 1.3 thus making the response string aberrant.

Table 4.4.6 below gives the minimum values of Δ_{ni} for which the contribution of the unexpected response to the outfit will exceed 1.2 and 1.3 for various test lengths.

Table 4.4.6 Minimum values of Δ_{ni} for various test lengths and two cut-off values for the outfit

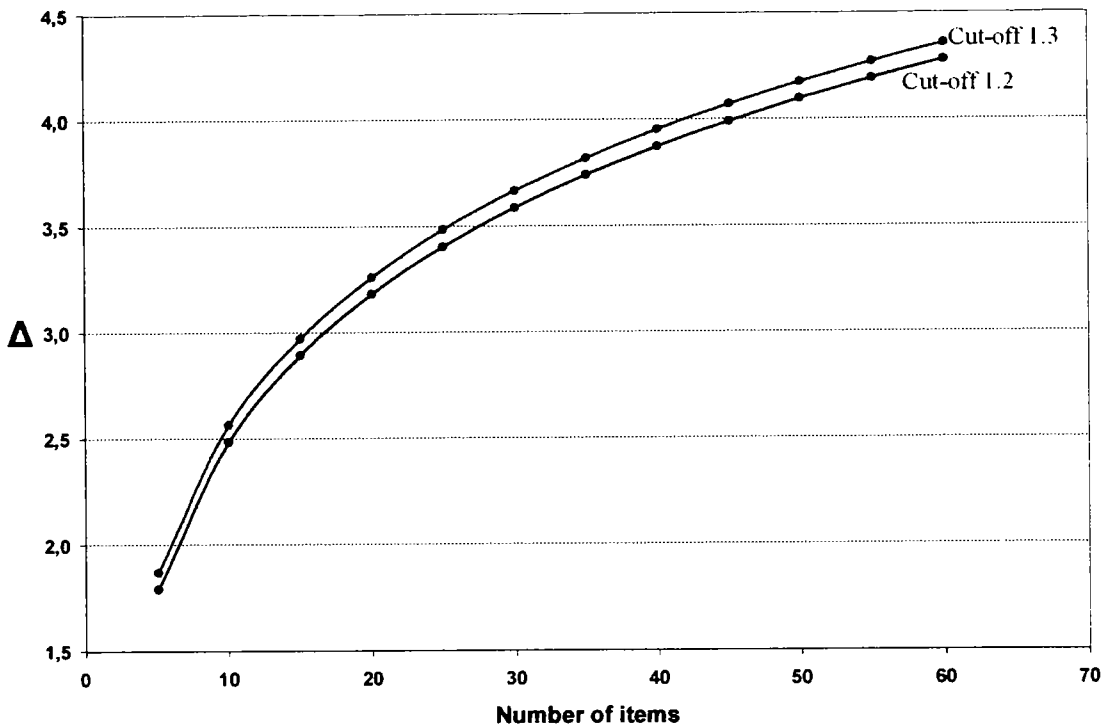
N	Minimum value of Δ_{ni}	
	Cut-off 1.2	Cut-off 1.3
5	1.792	1.872
10	2.485	2.565
15	2.890	2.970
20	3.178	3.258
25	3.401	3.481
30	3.584	3.664
35	3.738	3.818
40	3.871	3.951
45	3.989	4.069
50	4.094	4.174
55	4.190	4.270
60	4.277	4.357

One can see that, for example, in a 20-item test, one unexpected response on an item 3.178 logits away (above or below) from the ability estimate of a student would contribute 1.2 to the outfit, thus making the whole response string aberrant (in the case where 1.2 is used as the cut-off value for the outfit).

Similarly, in a 30-item test, and with 1.3 as the cut-off value, an unexpected response on an item with difficulty 3.664 logits away (above or below) from the ability estimate of the student would make the outfit exceed the cut-off value.

Figure 4.4.5 shows the same results in a diagrammatical form.

Figure 4.4.5 Minimum values of Δ_{ni} for various test lengths and two cut-off values for the outfit



On this diagram one can find the minimum value of Δ_{ni} for any test length up to 60 items and for an outfit cut-off value of 1.2 or 1.3.

With the formula $\Delta_{ni} = \ln(N \cdot C)$ one can find the minimum value of Δ_{ni} for any test length and any cut-off value for the outfit.

4.4.3 Investigating the effect of unexpected responses on the Outfit for items

Rasch measurement is not only used for the estimation of persons’ abilities or positions on the latent trait line. Another very common use is the validation of psychometric scales.

In such investigations the values of the infit and outfit are particularly important and play a key role in the assessment of the quality of the items.

The outfit is calculated in exactly the same way for persons and items. For persons, however the average of the squared standardised residuals is over all items and for the items over all persons.

Similarly, the contribution of any unexpected response to the item outfit is given by:

$$C = \frac{\text{Squared st. residual}}{\text{Sample size } n} \text{ and } C = \frac{e^{\Delta_{ni}}}{n} .$$

To show the effect of unexpected responses to the item outfit the researcher took an item of difficulty – 2 and a sample of persons (of various sizes) uniformly spread in the range from – 3.0 to 3.0 logits. Amongst these persons only two (with ability estimate of 2.0 and 2.5 logits, that is, 4 and 4.5 logits away from the item difficulty respectively) had unexpected wrong responses. All other persons responded exactly as expected (i.e. gave the most probable answer).

The researcher calculated the outfit values for sample sizes of 31, 41, 61, 77, 101, 121 and 151.

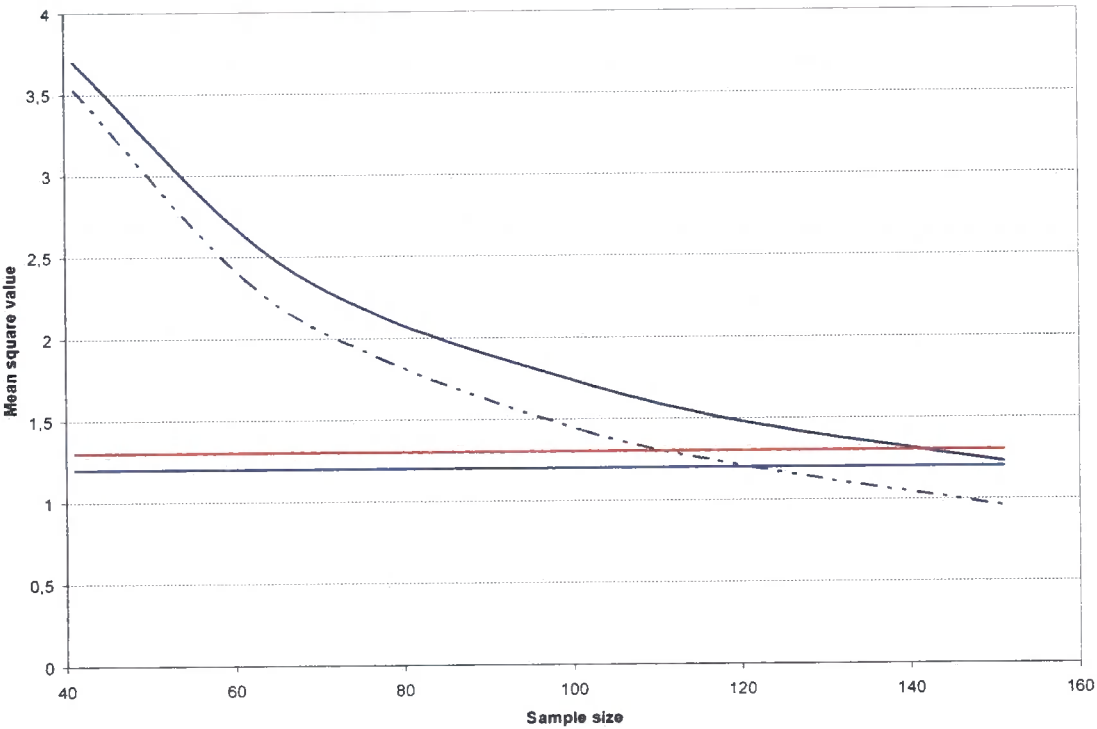
Table 4.4.7 below shows the contribution of the two unexpected responses separately and combined to the outfit value and the overall outfit value in all the cases.

Table 4.4.7 Contribution of two unexpected responses to the item outfit

Sample size	Ability 4 logits away	Ability 4.5 logits away	Combined contribution	Overall outfit
31	1.76	2.90	4.66	4.93
41	1.33	2.20	3.53	3.70
61	0.89	1.48	2.37	2.64
77	0.71	1.17	1.88	2.14
101	0.54	0.89	1.43	1.72
121	0.45	0.74	1.19	1.47
151	0.36	0.60	0.96	1.23

Figure 4.4.6 below shows the outfit values, the contribution to the outfit of the two unexpected responses and the cut-off values of 1.3 and 1.2.

Figure 4.4.6 Effect of test length on outfit



Two unexpected responses one on an item 4 logits away from the item estimate and one on an item 4.5 logits away will contribute to the item outfit more than 1.3 for a sample of 111 persons or less. This is just the contribution of the two unexpected responses. The overall outfit value (for the whole sample) will exceed 1.3 for a sample of about 140 persons or less.

Also, the two unexpected responses will contribute to the item outfit more than 1.2 for a sample of 120 or less. The overall outfit value (for the whole sample) will exceed 1.2 for a sample of about 155 persons or less

One can realize the impact a few highly unexpected responses can have to the outfit. The formula derived for the contribution of an unexpected response to the outfit can be used for any distance between ability and difficulty and any sample size. For example:

An item would be characterized as misfitting if it is answered by 500 persons, of which:

- (a) 5 with ability estimates 5 logits away from the item difficulty answer unexpectedly (Contribution to outfit = 1.484) or
- (b) 3 with ability estimates 5 logits away from the item difficulty and 2 with ability estimates 4.5 logits away answer unexpectedly (Contribution to outfit = 1.25) or
- (c) 4 with ability estimates 5 logits away from the item difficulty answer unexpectedly (Contribution to outfit = 1.187).

Therefore, if one tries to assess the validity of a test by administering it to about 500 persons, 4 or 5 highly unexpected responses to an item would make its outfit value exceed the cut-off score thus characterising it as misfitting even if the remaining persons respond to it as expected. That will not necessarily mean that the item is not functioning as expected or that it is a 'problematic' item that needs to be removed. Outfit measures suggest that these items (with the 4 or 5 unexpected person responses) misfit. However, especially if outfit is above, but close, to the cut-off value the identification should be a signal that the item should be investigated. No such item should be removed without investigation.

4.4.4 The number of items with 'less likely' answers for which Infit exceeds the cut-off value

This is an investigation into how many items with 'less likely' answers would make the infit exceed the cut-off value of 1.3, thus categorizing the response string as aberrant. The researcher used the term 'less likely' instead of 'unexpected' since these answers are on targeted items and therefore they are not unexpected in the same sense (highly unlikely) as in the outfit investigation. They are just the answers that have less chance of occurring.

In particular, the researcher investigated 3 different tests with lengths 12, 16 and 27 items (same test lengths as the maths tests used in this study) and in each case found the number of items with 'less likely' answers needed to make infit exceed the cut-off value of 1.3.

Finally various response strings and their corresponding infit and outfit values are presented.

The 12-item test

For this investigation a theoretical set of 12 dichotomous items was used with their difficulties following a rectangular distribution in the range from -2.0 to 2.0 logits. A hypothetical student of ability 0 logits was used, first with a deterministic response string.

Ability 0 was chosen, simply because this was the mean of item difficulties and the student would be centrally located amongst the items and would have the maximum possible number of well-targeted items to choose from.

Infit and outfit mean squares were calculated.

Then, in the second step, the correct response to the item closest to the student ability, on the left of the ability estimate (that is, item difficulty smaller than student's ability by 0.18 logits and a score of 1 on the deterministic response string), was changed into a wrong response and a score of 0 .

In the third step, the wrong response to the item closest to the student ability on the right of the ability estimate (difficulty 0.18 logits above the ability), and symmetrical to the first item changed, was also changed but this time into a correct response and a score of 1.

In the fourth step the correct response to the item second closest to the student ability, on the left, was changed into a 0, followed by the fifth step where the wrong response to the item second closest to the ability on the right was changed again into a 1.

The procedure continued in the same manner, with alternating items on the sides below and above the student ability, and getting further and further away from it.

At the end of each step the infit and outfit mean squares were calculated, until the infit exceeded the cut-off value of 1.3.

Table 4.4.8 below shows these calculations.

The first column shows all the steps and the items removed in the procedure. The first row, starting with ‘No items’, refers to the deterministic response string.

The second column headed by ‘Distance’ shows how many logits below (with a minus sign) or above the student ability the difficulty of the item removed was.

Finally, the last two columns show the infit and outfit mean square values.

Table 4.4.8 Infit and outfit calculations (N = 12)

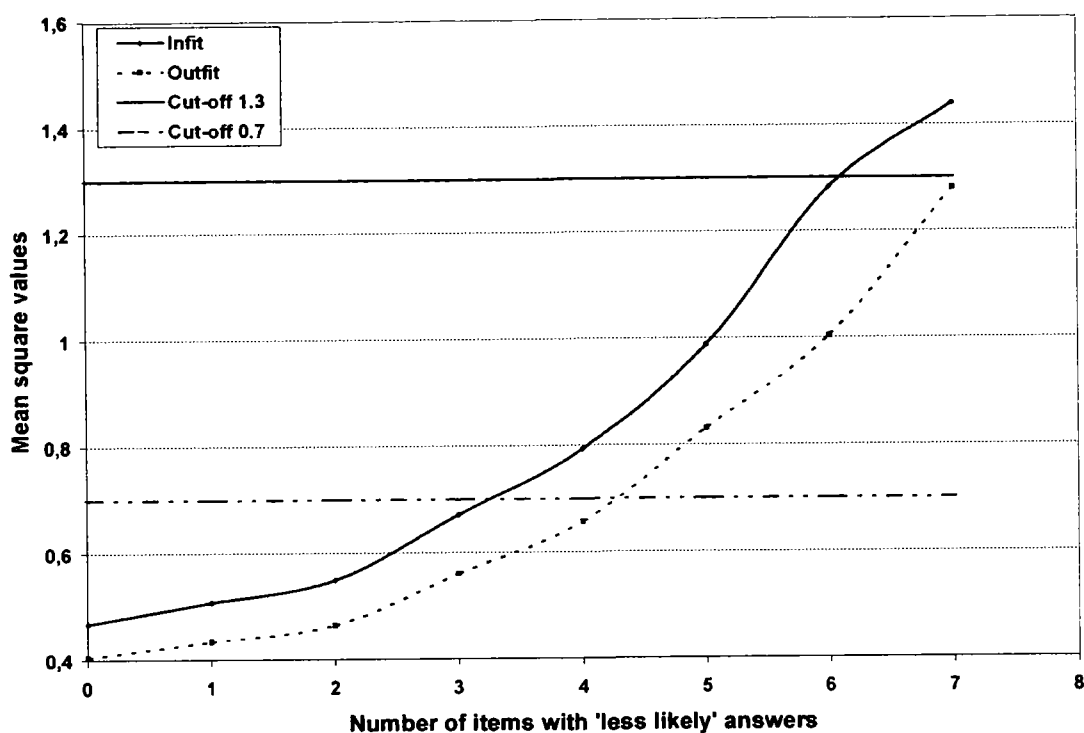
Steps	Distance	Infit	Outfit
No items		0.467	0.404
1 st item	- 0.18	0.508	0.434
2 nd item	0.18	0.549	0.464
3 rd item	- 0.55	0.671	0.560
4 th item	0.55	0.793	0.657
5 th item	-0.91	0.987	0.830
6 th item	0.91	1.282	1.003
7 th item	- 1.28	1.439	1.280

Figure 4.4.7 shows the change on the infit and outfit values as the number of responses with less likely responses increases.

It is evident, that, as expected, less likely responses on items close to the ability affect the infit mean square more than the outfit. At the same time, the minimum number of less likely responses needed to categorise the response string as aberrant (infit > 1.3) is 7 (out of the 12 items included in the test).

Furthermore, if a response string contains up to and including 3 (out of the 12) less likely responses, both infit and outfit will not exceed 0.7 thus categorizing the response string as overfitting.

Figure 4.4.7 the effect of less likely responses on infit and outfit ($N = 12$)



The response string below is the one which made the infit exceed the cut-off value. It is the first response strings (with the fewest less likely responses) that has been characterized by infit as aberrant. The dot in the middle of the string is where the ability of the person is located with respect to the items

Response string: 110000 • 111000

The 16-item test

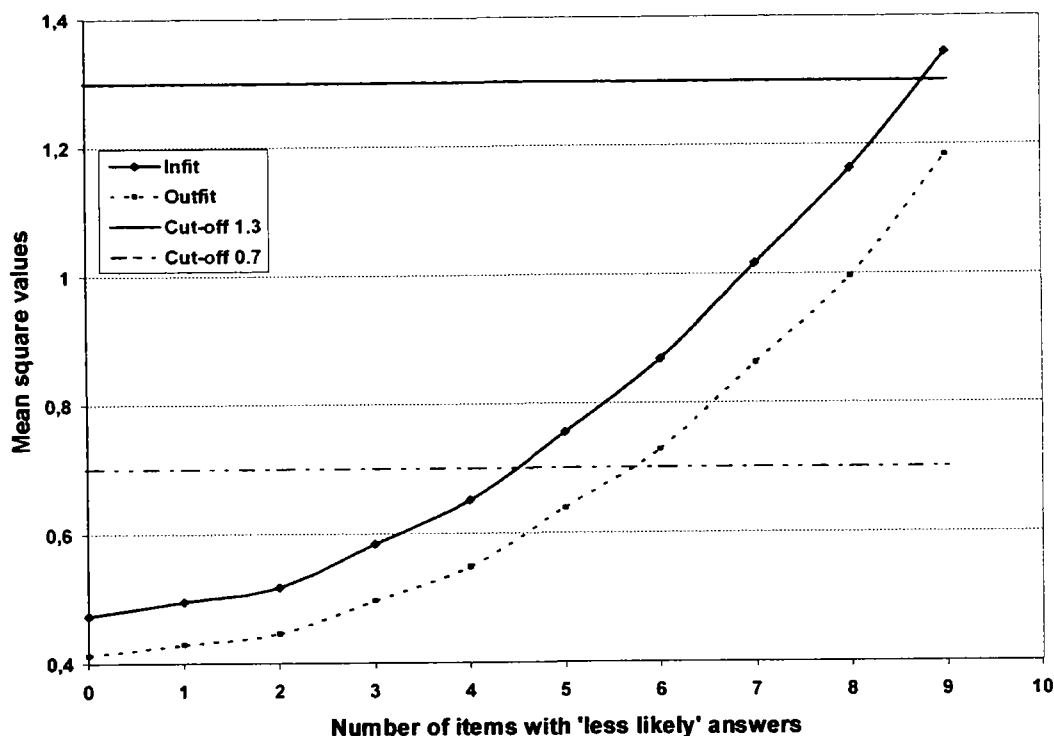
For this investigation the 16 dichotomous items were evenly spread in the range from – 2.0 to 2.0 logits. Again a hypothetical student of ability 0 logits was used, first with a deterministic response string. Then the procedure followed was identical to the one before by changing the more likely responses to the less likely ones, starting from the items closest to the student’s ability and infit and outfit statistics calculated at each step.

Table 4.4.9 shows all the calculations and figure 4.4.8 shows the effect on infit and outfit as the number of less likely responses increases.

Table 4.4.9 Infit and outfit calculations (N = 16)

Steps	Distance	Infit	Outfit
No items		0.473	0.412
1 st item	- 0.13	0.495	0.429
2 nd item	0.13	0.517	0.445
3 rd item	- 0.40	0.584	0.496
4 th item	0.40	0.651	0.548
5 th item	-0.67	0.756	0.638
6 th item	0.67	0.869	0.728
7 th item	- 0.93	1.016	0.862
8 th item	0.93	1.163	0.995
9 th item	-1.20	1.344	1.184

Figure 4.4.8 the effect of less likely responses on infit and outfit ($N = 16$)



It is evident again, that less likely responses on items close to the ability affect the infit mean square much more than the outfit. At the same time, the number of less likely responses needed to categorise the response string as aberrant ($\text{infit} > 1.3$) is 9 (out of the 16 items).

Furthermore, if a response string contains up to and including 4 less likely responses, both infit and outfit will not exceed 0.7 thus categorizing the response string as overfitting.

The response string below is the one which made the infit exceed the cut-off value. It is the first response strings (with the fewest less likely responses) that has been characterized by infit as aberrant.

Response string: 11100000 • 11110000

At first sight this response string looks very non-fitting; however these are not unexpected responses (i.e. not very unlikely responses). They are just the less likely responses and in particular the couple of items left and right of the ability that have almost the same probability (0.50) to get it right or wrong.

The 27-item test

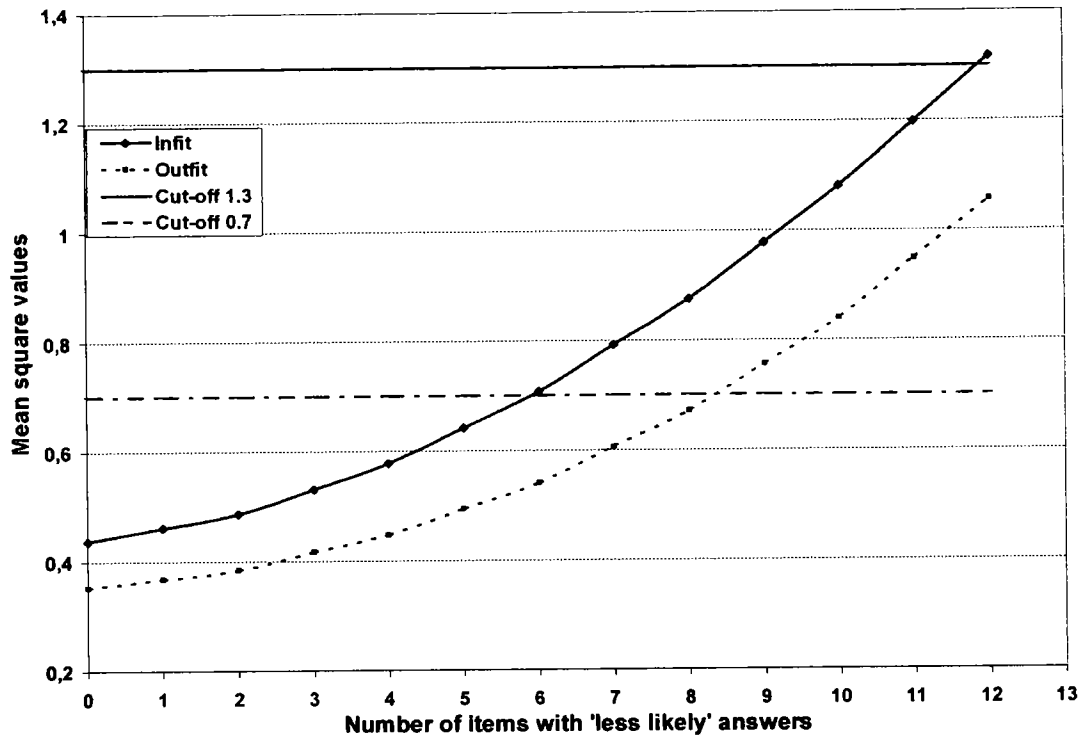
The same procedure was followed in this case too with one difference. The 27 items were spread evenly in the range from – 2.50 to 2.50 logits (because of the large number of items).

Table 4.4.10 below shows the infit and outfit calculations after each step, and figure 4.4.9 the effect on infit and outfit as the number of less likely responses increases.

Table 4.4.10 Infit and outfit calculations (N = 27)

Steps	Distance	Infit	Outfit
No items		0.437	0.352
1 st item	- 0.22	0.461	0.368
2 nd item	0.22	0.486	0.384
3 rd item	- 0.41	0.531	0.416
4 th item	0.41	0.577	0.447
5 th item	-0.60	0.642	0.494
6 th item	0.60	0.707	0.541
7 th item	- 0.79	0.792	0.606
8 th item	0.79	0.876	0.671
9 th item	-0.98	0.978	0.756
10 th item	0.98	1.080	0.840
11 th item	-1.17	1.198	0.948
12 th item	1.17	1.316	1.056

Figure 4.4.9 the effect of less likely responses on infit and outfit ($N = 27$)



In this case the minimum number of less likely responses needed to categorise the response string as aberrant ($\text{infit} > 1.3$) is 12 (out of the 27 items). Furthermore, if a response string contains up to and including about 6 less likely responses, both infit and outfit will not exceed 0.7 thus categorizing the response string as overfitting. Again, one can see that the effect of less likely responses to items close to the ability is larger on infit.

The response string below is the first response strings (with the fewest less likely responses) that has been characterized by infit as aberrant.

The score of 1 with a hat represents the point on the scale where the ability of the student is located. At that point, there is an item with difficulty 0. Therefore, for that item the probability of a correct response is the same as the probability of a wrong response ($= 0.5$). If the response to that item is 0 or 1 there is no difference on the infit or outfit calculations.

Response string: 1111111000000 $\hat{1}$ 11111110000000

Various response strings and their infit and outfit mean square values

Following Linacre’s and Wright’s (1994) example, the researcher then constructed table 4.4.11 with various response strings and their infit and outfit values. The number of items used was 20, with difficulties evenly spread in the range from – 2.0 to 2.0 logits and a mean of 0.

The student had a hypothetical ability of 0 thus located centrally amongst the items.

All unexpected responses in the table are shown in bold and all ‘less likely’ responses are underlined.

To make this investigation as close to real data as possible, all the response strings contained 10 correct and 10 wrong answers thus making the ability estimate very close to the mean item difficulty, since the ability is estimated using

$$\ln\left(\frac{\text{number of correct answers}}{\text{number of wrong answers}}\right) \text{ which gives } \ln\left(\frac{10}{10}\right) = \ln(1) = 0.$$

Finally, 0.7 and 1.3 were used as the cut-off values. Any values of infit or outfit below 0.7 indicate overfit and above 1.3 misfit.

The last column in the table gives a comment on each response string.

Table 4.4.11 Response strings and their mean-square fit statistics

	Response strings	Outfit	Infit	Comment
1	11111111110000000000	0.42	0.47	Deterministic (Guttman)
2	00 1111111111000000010	1.37	1.08	Misfit, high outfit (3 very unexpected answers)
3	11111111 <u>000</u> 101000000	0.57	0.67	Overfitting
4	111 <u>0</u> 11 <u>0</u> 11 <u>00</u> 101010000	0.95	1.02	Ideal (Based on Rasch model expectations)
5	1111 <u>0</u> 1 <u>0</u> 11 <u>0000</u> 1110000	0.99	1.08	Ideal (Based on Rasch model expectations)
6	111101 <u>00000</u> 111101000	1.20	1.33	Misfit, high infit (weird response string)
7	101 <u>0</u> 1010101010101010	1.72	1.59	Misfit, both high (student repeating a pattern)
8	111 <u>00000</u> 111101101000	1.41	1.49	Both high, (student may have missed a page)
9	0000000000 1111111111	3.24	2.91	Both high, too weird (possible miscoding)

The first response string (RS1) is a deterministic response string, and both infit and outfit are very low, much lower than 0.7.

RS2 has 3 unexpected responses on items with difficulties -2.0, -1.79 and 1.79 logits away from the ability estimate, and that shows on the high outfit value. The corresponding probabilities of occurrence of these responses are approximately 0.12, 0.14 and 0.14 respectively.

RS3 has 4 less likely answers but close to the ability estimate and is still overfitting the model.

RS4 and RS 5 are both ideal (as expected by the Rasch model, with infit and outfit values very close to 1). Both have 6 less likely answers.

RS6 is an unusual response string with 10 less likely answers, thus identified by the infit statistic as aberrant.

RS7, RS8 and RS9 all have high infit and outfit values. That is, all three are identified as aberrant by both fit statistics. All unexpected responses ($p < 0.15$) are in bold and all less likely responses are underlined. The comment in the last column gives a possible explanation about the misfit.

The table shows what was concluded earlier. Just a very few unexpected responses would make the outfit exceed the cut-off score, like the second response string on the table whereas for the infit to exceed 1.3 a much larger number of 'less likely' responses are needed.

In RS3 there are 3 unexpected responses, which is more than the number required to make outfit > 1.3 as shown in the outfit investigation (only 1 item was required) at the beginning of this chapter.

The explanation for this is quite simple: The unexpected responses in this case are on items whose difficulties are less than 2 logits away from the ability location (probability of occurring is 0.12 or more). In the outfit investigation the distance of the item difficulty used from the student's ability was 3 logits (probability of occurring is 0.047) and 4 logits (probability of occurring is 0.01).

These investigations (on outfit and infit) show the differences in the use and utility of the two mean-square statistics.

Outfit is ideal for detecting response strings with a few highly unexpected responses whereas, infit for detecting rather unexpected (or weird) response patterns, especially on items closer to the ability of the person's taking the test.

Concluding one can say that outfit can identify specific unexpected responses on items whose difficulty estimates is at some distance from the ability estimates and infit unexpected response patterns.

CHAPTER 5: CONCLUSIONS

This final chapter describes briefly the procedure followed in this study and discusses the findings with respect to the investigation:

- *into possible factors affecting misfit*
- *of reasons for students' unexpected responses, as described by the students themselves through interviews.*
- *as to whether misfit is an inherent characteristic of students.*
- *of predictive validity and internal consistency of raw scores of misfitting students*
- *of the infit and outfit mean square statistics*

5.1 The procedure

The main focus of this study was to explore the reasons behind aberrant response patterns in classroom mathematics tests.

The study took place in high schools (lyceums) in Cyprus and the data collection part was spread over two academic years, thus naturally dividing the whole project into two phases.

The main concern of the researcher in the first phase was the investigation of possible factors associated with misfit in the mathematics tests.

The main concerns of the researcher in the second phase were to investigate:

- Whether maths self-esteem and test anxiety are associated with misfit in classroom maths tests.
- Other possible reasons leading to aberrant response patterns, through in depth interviews of highly misfitting students.
- Whether misfit is an inherent characteristic of some students.
- Whether the internal consistency and predictive validity of scores of misfitting students are of a lower degree than those of the scores of the fitting students.

- How unexpected responses affect the person infit and outfit mean square statistics.

Before the commencement of data collection the researcher asked for, in writing, and received a written permission from the Director of Secondary Education, at the Ministry of Education and Culture, to proceed with his work.

For the analyses of the data collected for 6 out of the 7 assessment instruments (excluding the ADHD scale where previously defined cut-scores were employed) the Rasch models were used, the Partial Credit Model for the tests and the Rating Scale Model for the psychometric scales.

For the identification of aberrant response patterns infit and outfit mean square statistics were used, with cut-off values of 1.3 for the maths tests and 1.5 for the scales, as suggested by Wright et al. (1994) and Bond and Fox (2001).

For establishing the degree of reliability of the assessment instruments Cronbach's alpha was used, together with student reliability, separation index and strata. The last 3 are standard in the WINSTEPS (the software used) output. A high degree of reliability was established.

Given the importance of verifying the degree of validity of the assessment instruments the researcher collected a large amount of evidence to support the hypothesis that the degree of validity of the instruments used was high.

Table 5.1.1 below shows the different validation studies undertaken for each of the assessment instruments used:

Table 5.1.1 Validity evidence collected for the various instruments used

Evidence collected	Phase 1			Phase 2			
	Test 1	TAI	ADHD	Diagnostic	MSES	TAI 2	Test 2
Cronbach's alpha	•	•	•	•	•	•	•
Student reliability, separation index, strata	•	•		•	•	•	•
Factor Analysis	•	•	•		•	•	
PCA of stand. residuals	•			•	•	•	•
Invariance plots	•			•			•
Content validity questionnaire	•						•
Correlations with other criteria	•	•	•	•	•	•	•
Descriptive statistics and comparisons with published analyses		•				•	
Comparisons of teacher's ratings			•				
Male – female comparisons		•	•		•	•	

The reason for using factor analysis in the two scales was to compare the results with published data.

All the evidence collected, much to the delight of the researcher, confirmed the high degree of validity of all the instruments used.

5.2 Factors investigated for possible association with misfit

Since most of the factors (variables) considered in this study, as possible explanations of misfit, were categorical, or could easily be transformed to categorical, Log-linear analysis was used to investigate possible associations of these factors with misfit. The results suggest the following:

Student gender: Frary and Giles (1980) reported that females show fewer aberrant response patterns than males. There was no evidence of any association between student gender and misfit in the present study.

Teacher gender: There was no evidence of any association between teacher gender and student misfit.

Item arrangement: Two different item orders were used, an easy-to-hard and a more random order. There was no evidence of higher proportions of misfit in either of the two orders.

Different schools: Harnisch and Linn (1981) reported very different fit indices in different schools and attributed this to a mismatch between school curriculum and test content. Also Petridou and Williams (2007) attribute the class effect they have found to a mismatch between school curriculum and test content.

In this study although different schools were used the curriculum was common for all. No differences in the proportions of misfitting students were found in the different schools, and if different schools are indeed a factor affecting misfit, then Harnisch's and Linn's, and Petridou's and Williams' explanation may be true. That is, if there is association between different schools and misfit, this difference could be attributed to the mismatch between school curriculum and test content.

However, mismatch between test content and schools' curriculum raises questions about the validity of the tests used.

Another possibility that could explain differences between schools (although such differences were not found in this study) is different schemes of work adopted by different teachers, and in particular teachers in the same school may teach things differently or in a different order than teachers in other schools. However, results

showed no associations between the interaction of schools and different teachers on misfit in the tests with the polytomous items.

Ability: Keeves and Masters (1999) expressed concerns about different ability ranges affecting misfit differently. Curtis (2004), referring to Li and Olejnik (1997), suggests that the concerns expressed by Keeves and Masters are not a matter of great concern. Petridou and Williams (2007) report higher proportions of misfit in high ability students, having used a maths test with dichotomously scored items in their study.

The findings of this study support Curtis's suggestion for tests containing multistep problems since no association between different ability levels and misfit were found for the maths test in phase 1 and the first maths test in phase 2.

However, when the same analyses were performed on the misfitting students in test 2 of phase 2, the test containing a large number of multiple choice items, significant associations ($p = 0.022$) were found between ability and misfit with high ability students having higher proportions of misfit.

A probable explanation for the contradicting results of this study is the different item formats used. Significant association between ability and misfit were found only in the test containing the dichotomous items (as in Petridou's and Williams' study) whereas in this study in the tests with multistep problems with partial credit awarding no associations were found.

The problem with dichotomous items, even with construct-response items where the student has to provide the answer and may have to do calculations and follow a certain method to find the final answer, is that the scoring is either 0 for the wrong answer or 1 for the correct answer. In such a case, 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.

Therefore, this may well be the reason why high ability students are found to have higher proportions of misfit when dichotomous items are used. A high ability student is more likely to give an unexpected wrong answer (for example through carelessness) than a low ability student to give an unexpected right answer.

Test anxiety was suggested as a possible factor associated with misfit by many authors (such as Harnisch and Linn 1981; Bracey and Rudner, 1992; Athanasou and Lamprianou, 2002).

The findings of this study showed no such association in the maths test in phase 1 and the first test in phase 2. Some associations ($p = 0.018$) were found between the interaction of gender and test anxiety with misfit, with males with medium anxiety levels and females with low anxiety levels exhibiting larger proportions of misfit.

This again was on the test with the dichotomous items and the sample used was rather smaller than in all other investigations (386 students).

Furthermore, the combined effect of test anxiety and ability on misfit was investigated showing no associations, contradicting the results reported by Bracey and Rudner (1992).

ADHD: Distraction is suggested in the literature as a possible factor leading to aberrance and ADHD is a factor that leads to distraction in the classroom. This study showed no association between ADHD and aberrant response behaviour.

Math self-concept: There was no evidence of any association between student gender, math self-concept and misfit.

Motivation: Lamprianou and Boyle (2004) argued that examinees with too little motivation may be more likely to produce aberrant response patterns and suggest using the number of unauthorized absences as an indication of atypical schooling or low motivation. Following their suggestion, the researcher used the number of unauthorized absences during the first term of the academic year as an indication of

atypical schooling. No association between atypical schooling and misfit in the maths tests was found.

Other factors considered were:

- **Language competency** (Language grades)
- **Interest in Maths** (teachers' ratings)
- **Private tuition in maths** (58.8 % of the students took private tuition)
- **Study time** (in minutes)
- **Class revision for the test** (in teaching periods).

No associations were found between any of the above factors and misfit.

The only factor that showed some association with misfit was **different teachers**, where the results were significant ($p = 0.027$).

5.3 Results of the interviews

The analyses of the interviews suggest the following:

Unexpected responses occur mainly because of carelessness amongst the high scorers, and some times through wrong guessing.

For the low scorers aberrance occurs for completely different reasons.

Prior knowledge is a very common reason. Ten out of the 19 unexpected correct responses were due to prior knowledge.

This prior knowledge seems to occur at three different levels:

- From student to student (which can be eliminated if all students take the test at the same time).
- From classroom teacher to student (through hints about test items during the lessons prior to the test).
- From private tutor to student (through the tutors ability to guess what questions may be included in the test).

Other reasons include:

- **Cheating** (from more knowledgeable students)
- **Special preference.**

Sometimes, a correct response on a difficult item may be identified as unexpected through an artifactual situation. The investigation showed that if a high ability student misses a few marks through carelessness or through any other reason, then his/her ability is underestimated, making a correct response on a difficult item seem unexpected too.

Concluding the investigation about the artifactual situation it is the researcher's conclusion that:

- Misfit data can, in some cases be misleading. (The need for the infit mean square statistic originated from this realisation.)
- A student who is able and answers a large number of items correctly is much more likely to be identified as misfitting (even with only one careless mistake on an easy item) than a less able student who gets a small range and number of items correctly.
- If the items are dichotomous the less able student is unlikely to get hard items unexpectedly correct in a well designed test. On the other hand, a more able student is likely to get very easy items unexpectedly wrong, so long as there are many items. This gives an asymmetry in the data which might lead researchers to come to odd conclusions.

5.4 Is misfit an inherent characteristic of students?

Following the conclusions of the log-linear analyses, where all factors investigated showed no associations with misfit, the researcher wondered if misfit was an inherent characteristic of examinees, in particular 15 year old students taking classroom maths tests or answering psychometric scales.

The question that had to be answered then was "Do, more or less, the same students misfit in the administration of two assessment instruments?"

To answer this question, comparisons between the numbers of fitting and misfitting students were made in:

- Two maths tests (the ones taken in phase 2).
- Two psychometric scales (TAI and MSES) and
- A maths test and the TAI (in phase 1)

The Chi-square tests (contingency tables) used showed no association between misfit in the first and misfit in the second assessment instrument (p-values were 0.542, 0.392 and 0.895 respectively).

Smith (1986) and Lamprianou (2005) suggested that an individual with an aberrant response pattern may exhibit such response behaviour in other testing situations. The findings of this study do not support their suggestion.

5.5 Predictive validity

For the investigation of the predictive validity of the raw scores of misfitting students the correlation coefficients of the scores of fitting and misfitting students with other criteria were compared with the aid of confidence intervals. Other criteria included a second maths test, the first term maths grade and the maths final exam marks of the students.

No significant differences were found in the degree of predictive validity between fitting and misfitting students. This finding supports the findings of Lamprianou (2005).

5.6 Internal consistency

In order to investigate possible differences in the internal consistency of the responses of misfitting students the researcher used Cronbach's alpha together with confidence intervals as suggested by Iacobucci and Duhachek (2003).

Analyses showed that in all three maths tests used in this study the internal consistency of the responses of misfitting students were significantly lower (as in Lamprianou, 2005) than that of fitting students.

Following that, a more detailed investigation was carried out on the effect of high infit and high outfit values on the internal consistency.

Comparisons were made between the alphas of the responses of 4 different groups of students: fitting students, misfitting students with high outfit, with high infit and with both high. The comparisons showed that large outfit values do not affect internal consistency. On the other hand, large infit values do lower internal consistency. (This result was significant in all three tests).

This effect of infit only on internal consistency has motivated the researcher into investigating further the two mean square statistics, the outfit and infit.

5.7 Investigation of Outfit and Infit

Outfit mean square statistic

This investigation focused on the effect of test length on the outfit statistic of a response string with only one unexpected response. The aim was to find the largest possible test length for which the response pattern would be flagged by outfit as aberrant by only one unexpected response.

Analyses showed that the contribution of the unexpected response to the final outfit value (C), by itself only, exceeds the cut-off value of:

- 1.3 if the test length (N) is 15 items or less and
- 1.2 if N is 16 items or less.

Further analyses showed that if the response was highly unexpected (difficulty 4 logits away from ability) then such a test containing 53 items or less would have an outfit value above 1.3 (or with 60 items or less would have an outfit value above 1.2) thus categorizing the whole response string as aberrant.

The contribution of the unexpected response to the final outfit value (C), by itself only, exceeds the cut-off value of:

- 1.3 if N is 42 items or less and
- 1.2 if N is 45 items or less.

Therefore, if a classroom maths test contains less than 42 items, as is the common case with classroom tests, then only one unexpected response, on an item with difficulty 4 logits away from the student's ability, would make by itself the response string aberrant labeling the student as misfitting.

Finally the researcher, starting with $C = \frac{\text{Squared st. residual}}{\text{Testlength } N}$, worked out, with simple algebraic steps, a formula for the contribution of any single response to the

outfit. This formula simplifies to:

$$C = \frac{e^{\Delta_{ni}}}{N}$$

where $\Delta_{ni} = |B_n - D_i|$, that is, the positive difference between ability and difficulty.

These analyses raise the question whether such a response string, which is labeled as aberrant because of a high outfit value caused by only one unexpected response, can be considered invalid.

Based on the Rasch model the probability of a high ability student giving the wrong answer to an item 3 logits below his/her ability is approximately 5% (0.047) and to an item 4 logits below his/her ability is approximately 2% (0.018). Therefore, these unexpected responses, even though improbable will occur. If a high ability student scores one mark less than expected, this will not affect his/her ability estimate by much, especially if the test carries many marks.

A similar investigation into the contribution of a few unexpected responses to the item outfit value showed that an item could be characterized as misfitting, even if only 4 or 5 out of a sample of 500 persons (about 1% or less) responded highly unexpectedly.

That will not necessarily mean that the item is not functioning as expected or that it is a 'problematic' item that needs to be removed. Therefore, caution should be exercised in the assessment of the quality of items using the outfit mean square statistic.

Infit mean square statistic

This investigation focused on the number of 'less likely' responses needed to make the infit exceed the cut-off value, thus labeling the response string misfitting. The term 'less likely' instead of unexpected was used since these responses are on targeted items and therefore are not unexpected in the same sense (highly unlikely responses) as in the outfit investigation.

Three sets of theoretical dichotomous items were used: one with 12 items, one with 16 items, in both cases the item difficulties were uniformly spread over the interval - 2.0 to 2.0 logits, and one with 27 items uniformly spread over the interval -2.5 to 2.5 logits. These test lengths were selected because these were the exact lengths of the tests used in this study.

Analyses showed that:

- For the 12-item test it would take 7 or more 'less likely' items (that is more than half) to make infit exceed the cut-off value of 1.3. At the same time, with up to 3 'less likely' answers both infit and outfit were below 0.7, categorizing the response string as overfitting.
- For the 16-item test it would take 9 or more 'less likely' items (that is more than half) to make infit exceed the cut-off value of 1.3. At the same time, with up to 4 'less likely' answers both infit and outfit were below 0.7, categorizing the response string as overfitting.
- For the 27-item test it would take 12 or more 'less likely' items to make infit exceed the cut-off value of 1.3. At the same time, with up to 6 'less likely' answers both infit and outfit were below 0.7, categorizing the response string as overfitting.

Finally the researcher constructed a table (table 4.4.11 in chapter 4.4.4) with various response strings with their corresponding infit and outfit values, together with a comment about each response string, like the table presented by Linacre and Wright (1994).

5.8 Limitations of the study

It is important to critically evaluate the results and the whole study. The present study has some limitations that need to be taken into account when considering the study and its contributions. These limitations can also be seen as a means of improvement of future studies under the same theme. These limitations include:

- The findings and conclusions of this study relate to:
 - o The target population of first form (Lyceum) Cypriot students, of ages 15-16.
 - o Classroom maths tests.
 - o Tests containing multistep problems with partial credit awarding for partially correct answers.

Any generalization of the conclusions of this study beyond the target population and beyond this specific type of maths tests is unsafe.

- For the qualitative analyses only a small number of students (21) were interviewed from one of the schools involved. Although the results of these analyses are indicative of reasons (as perceived by the students themselves) for unexpected responses, no claim can be made with respect to the generalization of these results.
- The interviews were conducted by the teacher/researcher and although attempts were made to safeguard against bias like leading questions in the interview some bias may have been introduced because the interviewer, although not a class teacher of the students interviewed, he nevertheless was a teacher in the same school as the students interviewed.

- The marking of the tests was undertaken (as it is customary in classroom assessment) by the classroom teachers. Although clear instructions and marking schemes were given to the maths teachers one can not dismiss any claims that teacher subjectivity and degree of strictness may have affected uniform marking.
- Most of the teachers that were asked to complete the ADHD scale for their students were not very familiar with psychometric scales and that may well explain the high proportions (higher than found in the literature) of students described as meeting the criteria for ADHD.

5.9 Final comments

The fact that misfit is apparently not an inherent characteristic of students, together with the fact that almost none of the factors considered showed any association with misfit (when multistep items with partial credit awarding were used), leads to the conclusion that although misfit does occur, it is random and not dependent on psychological or demographic characteristics of the test takers.

This conclusion contradicts the main references or suggestions of possible factors affecting misfit found in the literature. Possible explanations for this contradiction include:

- The test items were, in the majority, multistep mathematical problems with partial credit awarding for partial success instead of the usual dichotomous items found in the literature.
- The low stakes status of the tests.
- The administration procedure, with the familiar classroom setting and the familiar, being-comfortable-with teacher-administrator of the test.

Some associations were found between ability and misfit ($p = 0.022$), between the interaction of gender and test anxiety on misfit ($p = 0.018$) and between different teachers and misfit ($p = 0.027$), the first two only in the case where mostly

dichotomous items were used. However given the large number of statistical tests performed this could be attributed to Type I error. Indeed, given the large number of statistical tests it would not be inappropriate to set the p value for statistical significance in this study at a more stringent point. This would probably mean that no statistically significant relationships were found.

Nevertheless, these possible associations should be investigated further.

Unexpected wrong answers (based on the explanations of the students themselves) occurred mainly because of carelessness and some times through wrong guessing amongst the high scorers. The reasons given for the low scorers, for unexpected correct answers, include prior knowledge, cheating and special preference.

The investigation into outfit and infit carried out in this study gives an explanation into why high infit is considered more of a threat to measurement and to validity (Linacre, 2006).

Although no differences were found between the predictive validity of scores of fitting and misfitting students, there were differences in the degree of reliability.

The response strings with high infit values had a significantly lower degree of reliability than those of fitting students or students with high outfit values.

The reason is simply because it takes only one (or very few) unexpected responses to make the outfit exceed the cut-off value whereas it takes many more 'less likely' responses (even more than half the number of items in short classroom tests with at least up to 16 items) to make the infit exceed the cut-off value.

This significantly lower degree of reliability of misfitting response patterns (with high infit) affects the degree of validity, since high degree of reliability is necessary for valid inferences.

And this is the reason why high infit is considered more of a threat to validity than high outfit.

Researchers (such as Karabatsos, 2003; Reise & Flannery, 1996; Rudner, 1983) have argued that aberrant responses may lead to misleading score interpretations and consequently to invalid measurement. However, it is the conclusion of the researcher that students with response patterns with high outfit values should not be considered as being invalidly measured without further investigation. Outfit can easily be

distorted by a single unexpected answer, even in a long test, and that, does not imply an invalid estimate of student's ability.

Similarly, items with response patterns with high outfit values should not be considered as malfunctioning and removed without investigation into the reasons of the high outfit values.

Furthermore, high infit is not caused by a few unexpected responses but by a large number of 'less likely' responses. Therefore, since the number of 'less likely' responses needed to make the infit value exceed the cut-off score is large, such an aberrant response string is a cause of worry about the validity of any interpretation based on it.

References

Allen, M. J. & Yen W. M. (1979). *Introduction to Measurement Theory*. Illinois: Waveland Press, Inc.

American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement and Evaluation [NCME]. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

American Psychiatric Association, (1994). *Diagnostic and Statistical Manual of Mental Disorders*. Washington D.C.: American Psychiatric Association.

Anastasi, A., and Urbina, S. (1997). *Psychological Testing*, 7th ed., NJ: Prentice Hall.

Andersen E. B. (1997). The Rating Scale Model. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 67-84). New York: Springer-Verlag New York Inc.

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

Baker, F. B. (1985). *The basics of item response theory*. Portsmouth, NH: Heinemann.

Barkley, R. A. (2000). *Taking charge of ADHD: The complete authoritative guide for parents*. New York: The Guilford Press.

Barkley, R. A., & Murphy, K. R. (1998). *Attention-Deficit Hyperactivity Disorder: A clinical workbook*. New York: The Guilford Press.

Blaxter, L., Hughes, C & Tight, M. (2001). *How to research*. Buckingham: Open University Press.

- Bode, R. K. (2004). Partial Credit Model and Pivot Anchoring. In Smith, E. V and Smith, R. M. (Eds.), *Introduction to Rasch Measurement* (pp. 279-295). Minnesota: JAM Press.
- Bond, T. G., and Fox, C. M. (2001). *Applying the Rasch model: fundamental measurement in the social sciences*. New Jersey: Lawrence Erlbaum Associates.
- Bond, T. G., and Fox, C. M. (2007). *Applying the Rasch model: fundamental measurement in the social sciences. 2nd Edition*. New Jersey: Lawrence Erlbaum Associates.
- Bracey, G., and Rudner, L. M. (1992). Person fit statistics: High potential and many unanswered questions. *Practical Assessment Research and Evaluation* 3(7). Retrieved August 24, 2004 from <http://pareonline.net/getvn.asp?v=3&n=7>.
- Brewer, J., and Hunter, A. (1989). *Multimethod research: A synthesis of styles*. Newbury Park, CA: Sage
- British educational Research Association [BERA] (2000). *Good practice in educational research writing*. Retrieved October 16, 2007 from www.bera.ac.uk/publications/pdfs/GOODPR1.PDF
- Brown, S. (1997). Respondents' comments. In S. Hegarty (Ed.) *The role of research in mature education systems*. (pp. 81-88), Slough: National Foundation of Educational Research.
- Callingham, R. and Watson, J. M. (2005). Measuring Statistical Literacy. *Journal of Applied Measurement*, 6(1), 19-47.
- Campbell, D. T. (1969). Reforms as experiments. *American psychologist*, 24, 409-429.
- Campbell, D. T. (1974). *Qualitative knowing in action research*. Paper presented at the annual meeting of the American Psychological Association. L.A. California.

- Chen, S. P. C., Bezruczko, N., and Ryan-Henry (2006). Rasch analysis of a new construct: Functional caregiving for adult children with intellectual disabilities. *Journal of Applied Measurement*, 7(2), 141-159.
- Creswell, J. W. (2003). *Research design: Qualitative, quantitative and mixed methods approaches*. (2nd ed.). Thousand Oaks, CA: Sage Publications Inc.
- Cronbach, L. J. (1974). *Beyond the two disciplines of scientific psychology*. Paper presented at the annual meeting of the American Psychological Association. L.A. California.
- Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.
- Crocker, L. & Algina, J. (1986) *Introduction to Classical and Modern Test Theory*. Orlando: Harcourt, Brace, Jovanovich College Publishers.
- Curtis, D. D. (2004). Person Misfit in Attitude Surveys: Influences, Impacts and Implications. *International Educational Journal*, 5(2), 125-144.
- Derver, E. (1997). *Using semi-structured interviews in small-scale research: A teacher's guide*. Glasgow: SCRE Publications.
- De Landsheere, G. (1993). History of Educational Research. In M. Hammersley (Ed.), *Educational research: Current issues*. London: Paul Chapman Publishing Ltd.
- Dickson, H. G. & Kohler, F. (1996). The multi-dimensionality of the FIM motor items precludes an interval scaling using Rasch analysis. *Scandinavian Journal of Rehabilitation Medicine*, 26, 159-162.
- Douglas, G. A. (1990). Response patterns and their probabilities. *Rasch Measurement Transactions*, 3(4), 75-77. Retrieved August 12, 2004 from <http://www.rasch.org/rmt/rmt34a.htm>.

- Drasgow, F., Levine, M. V., and Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardised indices. *British Journal of Mathematical and Statistical Psychology*, 38, 67-86.
- Dubois, P. H. (1970). *A history of Psychological Testing*. Boston: Allyn and Bacon
- Duhachek, A. and Iacobucci, D. (2004). Alpha's standard error (ASE): An accurate and precise confidence interval estimate. *Journal of Applied Psychology*, 89 (5), 792 – 808.
- Emons, W. H. M., Glas, C. A. W., Meijer, R. R., and Sijtsma, K. (2003). Person-Fit in Order-Restricted Latent Class Models. *Applied Psychological Measurement*, 27(6) 459-478.
- Fischer, G. H. and Molenaar, I. W (1995). *Rasch Models: Foundations, Recent developments and Applications*. New York: Springer-Verlag New York Inc.
- Frary, R. B. (1982, March). *A comparison of person fit measures*. Paper presented at the annual meeting of the American Educational Research Association, New York, USA.
- Frary, R. B., and Giles M. B. (1980). *Multiple-choice test bias due to answering strategy variation*. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, Massachusetts.
- Glas, C. A. W., and Meijer, R. R. (2003). A Bayesian approach to Person Fit analysis in Item Response Theory models. *Applied Psychological Measurement*, 27(3) 217-233.
- Goldstein, H. (1979). Consequences of using the Rasch model for educational assessment. *British Educational Research Journal*, 5(2), 211-220.

- Goldstein, H. (2004). *The education world cup: international comparisons of student achievement*. Plenary talk to Association for Educational Assessment – Europe: Budapest, Nov. 4 – 6, 2004.
- Hambleton, R. K. (1993). Principles and selected applications of Item Response Theory. In Linn, R. L. (Eds.), *Educational Measurement*. (3rd ed.), (pp 13-104), Phoenix: Oryx Press.
- Hambleton, R. K. Swaminathan, H. Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. California: SAGE Publications, Inc.
- Hargreaves, D. H. (1967). *Social relations in a secondary school*. London: Routledge and Kegan Paul.
- Harnisch, D. L., and Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement* 18(3), 133-146.
- Higher Education Funding Council of England [HEFCE], (1999). *Research assessment exercise 2001: consultation on assessment panels' criteria and working methods*. Bristol: HEFCE.
- Hillage, L., Pearson, R., and Tamkin, P. (1998). *Excellence in Research in Schools*. Brighton: Institute of employment studies.
- Howe, K. R. (1988). Against the quantitative-qualitative incompatibility thesis or dogmas die hard. *Educational Researcher*, 17, 10-16.
- Howell, D. C. (1992). *Statistical Methods for Psychology*. USA: Pw S-Kent Pub. Co
- Iacobucci, D. & Duhachek, A. (2003). Advancing alpha: Measuring reliability with confidence. *Journal of Consumer Psychology*, 13(4), 478-487.

- Institute for Objective Measurement, (2000). Definition of Objective Measurement. Retrieved January 22, 2008 from <http://www.rasch.org/define.htm>.
- Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement*, 1(2), 152-176.
- Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty six person-fit statistics. *Applied Measurement in Education*, 16(4), 277-298.
- Keeves, J. P., & Alagumalai, S. (1999). New approaches to measurement. In G. N. Masters and J. P. Keeves (Eds), *Advances in measurement in educational research and assessment* (pp. 23-42). Amsterdam: Pergamon.
- Kerlinger, F. N. (1986). *Foundations of Behavioral Research*, (3rd Edition). Orlando: Holt, Rinehart and Winston, Inc.
- Klauer, K. C. (1995). The assessment of person fit. In G. H. Fischer and I. W. Molenaar (Eds), *Rash models: Foundations, recent developments and applications* (pp. 97-110). New York: Academic Press.
- Kline, P. (2000). *A psychometrics primer*. London: Free Association Books.
- Kline, P. (1994). *An Easy Guide to Factor Analysis*. London: Routledge
- Koning, A. J., and Franses, P. H. (2003). Confidence intervals for Cronbach's coefficient alpha values. ERIM Report Series Research in Management, ERS – 2003 – 041 – MKT.
- Lamprianou, I. (2002). *Optimal appropriateness measurement in the context of the one parameter logistic model*. Unpublished thesis. University of Manchester.
- 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, Montreal, Canada.

Lamprianou, I. (2006). The stability of marker characteristics across tests of the same subject and across subjects. *Journal of Applied Measurement*, 7(2), 195 – 205.

Lamprianou, I., and Boyle, B. (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(3), 239 – 259.

Lee, N. P., and Fischer Jr, W. P. (2005). Evaluation of the Diabetes Self-Care Scale. *Journal of Applied Measurement*, 6(4), 366 – 381.

Levine, M. V. & Drasgow, F. (1988). Optimal Appropriateness Measurement. *Psychometrika*, 53(2), 161-176.

Levine, M. V. and Rubin, D. B. (1979). Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics*, 4, 269 – 290.

Linacre, J. M. (1996). The Rasch Model cannot be “Disproved”. *Rasch Measurement Transactions*, 10(3) 512-514.

Linacre, J. M. (1998a). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement*, 2(3), 266-283.

Linacre, J. M. (1998b). Structure in Rasch residuals: Why principal components analysis? *Rasch Measurement Transactions*, 12(2) 636.

Linacre, J. M. (1999). Relating Cronbach and Rasch Reliabilities. *Rasch Measurement Transactions*, 13(2), 696. Retrieved December 18, 2004, from <http://www.rasch.org/rmt/rmt132i.htm>

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

- Linacre, J. M. (2006). WINSTEPS (3.61.2) [Computer Software] Chicago: Winsteps.com.
- Linacre, J. M., and Wright, B. D. (1994). Chi-square Fit Statistics. *Rasch Measurement Transactions*, 8(2), 360. Retrieved July 18, 2007 from <http://www.rasch.org/rmt/rmt82a.htm>
- Lincoln, Y. S., and Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.
- Lyman, H. B. (1998). *Test scores and what they mean*. (6th ed.). Boston: Allyn and Bacon.
- MacGaw (1997) The nature and function of educational research. In S. Hegarty (Ed.) *The role of research in mature education systems*, (pp. 59-80), Slough: National Foundation of Educational Research.
- Marsh, H. W., Byrne, B. M., and Shavelson, R. J. (1988). A multifaceted academic self-concept: its hierarchical structure and its relation to academic achievement. *Journal of Educational Psychology*, 80, 366-380.
- Marsh, H. W., and Shavelson, R. J. (1985). Self-concept: Its multifaceted hierarchical structure. *Educational Psychologist*, 20, 107-125.
- Marsh, H. W., and O'Neal, R. (1984). Self Description Questionnaire III: The construct validity of multidimensional self-concept ratings by late adolescents. *Journal of Educational Measurement*, 21(2), 153-174.
- Massof, R. W. & Fletcher, D. C. (2001). Evaluation of the NEI visual functioning questionnaire as an interval measure of visual ability in low vision. *Vision Research*, 41(3), 397 – 413.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.

- Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement*, 25 (1), 15-29.
- Masters, G. N. (1993). Undesirable item discrimination. *Rasch Measurement Transactions*, 7(2), 289. Retrieved January 12, 2008 from <http://www.rasch.org/rmt/rmt172f.htm>
- Masters, G. N. (2001). *The key to objective measurement*. Retrieved May 15, 2007 from www.rasch.org/kev.pdf
- Masters, G. N., and Keeves, J. P. (1999). *Advances in measurement in educational research and assessment*. Amsterdam: Pergamon.
- Masters, G. N., and Wright, B. D. (1997). The Partial Credit Model. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 101-121). New York: Springer-Verlag New York Inc.
- Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and practice*. 11(1), 3-9.
- Mehrens, W. A. & Lehmann, I. J. (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(1), 3-8.
- Meijer, R. R., Molenaar, I. W., and Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement*, 18(2), 111-120.
- Meijer, R. R. Sijtsma, K. (2001). Person Fit Statistics: What is their purpose? *Rasch Measurement Transactions*. 15(2) 823. Retrieved May 12, 2007 from <http://www.rasch.org/rmt/rmt152d.htm>

- Merrell, C., & Tymms, P. (2001). Inattention, hyperactivity and impulsiveness: Their impact on academic achievement and progress. *British Journal of Educational Psychology*, 71, 43-56.
- Merrell, C., & Tymms, P. (2005). Rasch analysis of inattentive, hyperactive and impulsive behaviour in young children and the link with academic achievement. *Journal of Applied Measurement*, 6(1), 1-8.
- Messick, S. (1993). Validity. In Linn, R. L. (Eds.), *Educational Measurement*. (3rd ed.). pp 13-104. Phoenix: Oryx Press.
- Messick, S. (1995). Standards of Validity and the Validity of Standards in Performance Assessment. *Educational Measurement: Issues and Practice*, 14(4), 5 – 8.
- Michell, J. (2003). Measurement: A beginner's guide. *Journal of Applied Measurement*, 4(4), 298-308.
- Mok, M. M. C. (2004). Validation of scores from self-learning scales for primary students using true-score and Rasch measurement methods. *Journal of Applied Measurement*. 5 (3), 258-286.
- Molenaar, I. W., and Hoijtink H. (1990). The many null distributions of person-fit indices. *Psychometrika*, 55(1), 75 – 105.
- Molenaar, I. W., and Hoijtink H. (1996). Person-fit and the Rasch model, with an application to knowledge of logical quantors. *Applied Measurement in Education*, 9(1), 27 – 45.
- Mortimore, P. (2000). Does Educational Research matter? *British Educational Research Journal*, 26(1), 5-24.

Myford, C. M., and Wolfe, E. W. (2002). When raters disagree, then what: Examining a third rating discrepancy resolution procedure and its utility for identifying unusual patterns of ratings. *Journal of Applied Measurement*, 3(3), 300 – 324.

Nunnally, J. C., and Bernstein, I. H. (1994) *Psychometric Theory*, (3rd ed.). USA: McGraw-Hill, Inc.

O'Reilly, R. P., and Wightman, L. E. (1971). Improving the identification of anxious elementary school children through the use of an adjusted anxiety scale. *Journal of Educational Measurement*. 8(2), 107-112.

Panayides, P., Merrell, C., and Tymms, P. (2008). *Does severely Inattentive, Hyperactive and Impulsive Behaviour lead students to aberrant response patterns in mathematics tests?* Retrieved May 28, 2008, from <http://www.cemcentre.org/documents/CEM%20Extra/SpecialInterests/ADHD/ADHD%20and%20Misfit%208.pdf>

Petridou, A. and Williams, J (2007). Accounting for aberrant test response patterns using multilevel models. *Journal of Educational Measurement* 44(3), 227 – 247.

Plake, B. S., Ansorge, C. J., Parker, C. S., and Lowry, S. R. (1982). Effects of Item Arrangement, Knowledge of Arrangement, Test Anxiety and Sex on Test Performance. *Journal of Educational Measurement*. 19(1), 49-57.

Prieto, L., Roset, M., and Badia, X. (2001). Rasch measurement in the Assessment of Growth hormone Deficiency in adult patients. *Journal of Applied Measurement*, 2(1), 48 – 64.

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

Raiche, G. (2005). Critical eigenvalue sizes in standardized residual principal components analysis. *Rasch Measurement Transactions*. 19(1) 1012.

- Reichardt, C. S., and Rallis, S. F. (1994). Qualitative and quantitative inquiries are not incompatible. A call for a new partnership. In C. S. Reichardt and S. F. Rallis (Eds.), *The qualitative-quantitative debate: New perspectives*. (pp 85-92). San Francisco: Jossey-Bass.
- Reise, S. P. and Flannery, W. P. (1996). Assessing person-fit on measures of typical performance. *Applied Measurement in Education*, 9(1), 9 – 2.
- Rice, F. P. (1999). *The adolescence: Development, relationships and culture*. London: Allyn and Bacon.
- Roberts, G. (2003). Review of research assessment (Report to the UK funding bodies). Retrieved March 12, 2008 from <http://www.rareview.ac.uk/reports/roberts.asp>
- Rudman, H. C. (1989). Integrating testing with teaching. *Practical Assessment, Research & Evaluation*, 1(6). Retrieved from <http://PAREonline.net/getvn.asp?v=1&n=6>.
- Rudner, L. M. (1983). Individual Assessment Accuracy. *Journal of Educational Measurement*, 20(3), 207-219.
- Sarason, I. G., and Palola, E. G. (1960). The relationship of test and general anxiety, difficulty of task, and experimental instructions to performance. *Journal of Experimental Psychology*, 59, 185-191.
- Schumacker, R. E., and Linacre, J. M. (1996). Factor analysis and Rasch. *Rasch measurement transactions*, 9(4), 470. Retrieved from <http://rasch.org/rmt/rmt94k.htm>
- Shavelson, R. J. (1988). Contribution of educational research to policy and practice. *Educational Researcher*, 17, 2-4.
- Shavelson, R. J., Hubner, J. J., and Stanton, G. C. (1976). Self-concept validation of construct interpretations. *Review of Educational Research*, 46, 407-441.

- Skaalvik, S. & Skaalvik, E. M. (2004). Gender differences in math and verbal self-concept, performance expectations, and motivation. *Sex Roles: A Journal of Research*. Retrieved from http://www.findarticles.com/p/articles/mi_m2294/is_3-4_50/ai_114703694.
- Smith, J. K., and Heshusius, L. (1986). Closing down the conversation: The end of the quantitative-qualitative debate among educational researchers. *Educational Researcher*, 15, 4-12.
- 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(4), 78-79. Retrieved August 12, 2004, from <http://www.rasch.org/rmt/rmt34b.htm>
- Smith, R. M. (1992). *Applications of Rasch Measurement*. Minnesota: Jam Press.
- Smith, R. M. (1996). Polytomous Mean-Square Fit Statistics. *Rasch Measurement Transactions*. 10(3). 516-517. Retrieved July 29, 2008 from <http://www.rasch.org/rmt/rmt103a.htm>
- Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement*, 1(2), 199-218.
- Smith, Jr., E. V. (2000). Understanding Rasch measurement: Metric development and score reporting in Rasch measurement. *Journal of Applied Measurement*, 1(3), 303 – 326.
- Smith, Jr., E. V. (2004a). Evidence for the Reliability of Measures and Validity of Measure Interpretations: A Rasch Measurement Perspective. In E. V Smith Jr, and R. M. Smith (Eds), *Introduction to Rasch Measurement* (pp. 93-122). Minnesota: JAM Press.

Smith, Jr., E. V. (2004b). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal components analysis of residuals. In E. V Smith Jr, and R. M Smith (Eds), *Introduction to Rasch Measurement* (pp. 575-599). Minnesota: JAM Press.

Spielberger, C. D. (1980). *Test Anxiety Inventory: Preliminary Professional Manual*. USA: Mind Garden.

Stanley, G. V. (1991). On the importance of educational research. (1990 WAIER Research Forum Opening Address). *Issues in Educational Research*, 1(1), v-vi.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103, 677-680.

Stone, M. H. (2004). Substantive Scale Construction. In E. V. Smith Jr, and R. M. Smith, (Eds.), *Introduction to Rasch Measurement* (pp. 201-225). Minnesota: JAM Press.

Stout, W.F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52 (4), 589-617.

Tasker, M., and Packham, D. (1998, October 8). "Research shows" Or does it? *Independent*. Retrieved July 3, 2008 from <http://people.bath.ac.uk/mssdep/810ind.htm>

Tashakkori, A., and Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches*. London: Sage.

Thissen, D., and Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, 47, 397 – 412.

Thorndike, E.L. (1904). *An introduction to the theory of mental and social measurement*. New York: Teacher's College.

Thurnstone, L. L. (1931). Measurement of social attitudes. *Journal of Abnormal and Social Psychology*. 26, 249 – 269.

Towle, N. J., and Merrill, P. F. (1975). Effects of Anxiety Type and Item-Difficulty Sequencing on Mathematics Test Performance. *Journal of Educational Measurement*. 12 (4), 241-249.

Wainer, H. (1996). Depicting error. *The American Statistician*. 50(2), 101-111.

Western Australian Institute for Educational Research (1991). Strategic review of research in education. *Issues in educational research* 1(1), 43-45. Retrieved September 20, 2007 from <http://education.curtin.edu.au/iier1/waier-viewpoint.html>.

Westmarland, N. (2001) The quantitative/qualitative debate and feminist research: A subjective view of objectivity. *Forum: Qualitative Social Research*, 2(1). Retrieved May 30, 2007 from <http://qualitative-research.net/fqs-eng.htm>

Wikipedia (no date). *Research Assessment Exercise*. Retrieved March 12, from http://en.wikipedia.org/wiki/Research_Assessment_Exercise

Wright, B.D. (1967). *Sample-free Test Calibration and Person Measurement*. ETS Invitational Conference on Testing Problems. Retrieved May 24, 2004 from <http://www.rasch.org/memo1.htm>.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement* 14(2), 97-115.

Wright, B.D. (1983). *Fundamental measurement in social science and education*. Research Memorandum No. 33a. MESA Psychometric Laboratory.

Wright, B. D. (1988). Rasch model for Thurstone's scaling requirements. *Rasch Measurement Transactions*, 2(1) 13 – 14. Retrieved September 12, 2007, from <http://www.rasch.org/rmt/rmt21a.htm>.

Wright, B.D. (1993). Equitable Test Equating. *Rasch Measurement Transactions* 7(2) 298-299 Retrieved May 24, 2004 from <http://www.rasch.org/rmt/rmt72q.htm>.

Wright, B.D. (1995). 3PL or Rasch?. *Rasch Measurement Transactions* 9(1) 408 Retrieved from <http://www.rasch.org/rmt/rmt91b.htm>

Wright, B.D. (1997) *Measurement for social science and education: A history of social science measurement*. MESA, Memo 62. Retrieved October 24, 2006, from <http://www.rasch.org/memo62.htm>

Wright, B.D. (1999). *Fundamental measurement for psychology*. MESA Memo 64. Retrieved December 27, 2006 from <http://www.rasch.org/memo64.htm>

Wright, B. D. and Linacre, J. M. (1987). Rasch model derived from objectivity. *Rasch Measurement Transactions*, 1(1) 5-6. Retrieved from <http://www.rasch.org/rmt/rmt11a.htm>

Wright, B. D., and Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation*, 70, 857-860. Retrieved from <http://www.rasch.org/memo44.htm>

Wright, B. D., Linacre, J. M., Gustafson, J-E., and Martin-Lof, P. (1994). Reasonable mean square fit values. *Rasch measurement transactions*, 8(3), 370. Retrieved July 10, 2000, from <http://www.rasch.org/rmt/rmt83b.htm>

Wright, B. D., and Masters, G. N. (1982). *Rating Scale Analysis*. Chicago: Mesa

Wright, B. D., and Mok, M. M. C (2004). An Overview of the Family of Rasch Measurement Models. In E. V. Smith Jr., and R. M. Smith (Eds.), *Introduction to Rasch Measurement* (pp. 1-24) Minnesota: JAM Press.

Wright, B. D., and Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 29, 23-48.

Yates, L. (2002). *What does good educational research look like?* Paper presented at Why Learning? Seminar, Australian Museum/University of Technology, Sydney.

Appendices

1. Permission from the Head of Secondary Education at the Ministry of Education and Culture (In Greek).
2. Written consent from one of the Headmasters (In Greek).
3. Consent form given to teachers and students (In Greek).
4. Maths test in phase 1, on straight line graphs (Translated in English).
5. Maths diagnostic test in phase 2 (Translated in English).
6. Maths test in phase 2, on quadratic equations (Translated in English).
7. Test Anxiety Inventory (In Greek).
8. Test Anxiety Inventory (Back translation from Greek).
9. Maths Self- Esteem scale.
10. Maths Self- Esteem scale (In Greek).
11. ADHD scale (in Greek).
12. Questionnaire on Content Validity of the maths tests.
13. Questionnaire on Content Validity of the maths tests (In Greek).
14. Statistical Methods: Difference between two correlation coefficients and Log-linear analysis.
15. Log-linear analysis tables-results.

Appendix 1

Permission from the Head of Secondary Education at the Ministry of
Education and Culture (In Greek)

ΚΥΠΡΙΑΚΗ



ΔΗΜΟΚΡΑΤΙΑ

Αρ. Φακ.: 7.19.46.7/2

ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΠΟΛΙΤΙΣΜΟΥ
ΓΡΑΦΕΙΟ ΔΙΕΥΘΥΝΤΗ ΜΕΣΗΣ ΕΚΠΑΙΔΕΥΣΗΣ
1434 ΛΕΥΚΩΣΙΑ

3 Φεβρουαρίου, 2005

↓ Παναγιώτης Παναγίδης
Νίκου Καβαδία 1
Κ. Πολεμίδια
Λεμεσός 4152

**Θέμα: Διεξαγωγή Έρευνας στα πλαίσια ερευνητικής εργασίας
«Misfitting Students in Mathematics Tests»**

Αναφορικά με την επιστολή σας με ημερομηνία 14/1/2005, σχετικά με το πιο πάνω θέμα σας πληροφορώ ότι το Υπουργείο Παιδείας και Πολιτισμού εγκρίνει την αίτησή σας για διεξαγωγή έρευνας με δείγμα 500 μαθητών στα Λύκεια, Πολεμιδιών, Αγίου Νικολάου, Ιδαλίου και Αγίου Νεοφύτου, νοουμένου ότι:

- (α) Δε θα χαθούν μαθήματα
- (β) Θα εξασφαλιστεί προηγουμένως η συγκατάθεση των διευθυντών και των γονιών των μαθητών.
- (γ) Δε θα υπάρξουν οικονομικές απαιτήσεις από το Υπουργείο Παιδείας και Πολιτισμού και
- (δ) Τα αποτελέσματα θα κοινοποιηθούν τόσο στο Υπουργείο Παιδείας και Πολιτισμού όσο και στο Παιδαγωγικό Ινστιτούτο.

Ανδρέας Σκοτεινός
Διευθυντής Μέσης Εκπαίδευσης

Κοιν: Διευθυντή Παιδαγωγικού Ινστιτούτου

Appendix 2

Written consent from one of the Headmasters (In Greek)

ΔΗΛΩΣΗ ΣΥΓΚΑΤΑΘΕΣΗΣ ΔΙΕΥΘΥΝΤΗ

Ο υποφαινόμενος Μιχάλης Οικονομίδης, διευθυντής του Λυκείου Πολεμιδιών, συγκατατίθεμαι στη συμμετοχή μαθητών της Α' Λυκείου στην έρευνα του κυρίου Παναγιώτη Παναγίδη καθηγητή μαθηματικών στο Λύκειο Πολεμιδιών, με τίτλο "Misfitting students in Mathematics Tests" στα πλαίσια ερευνητικής εργασίας η οποία διεξάγεται μετά από έγκριση του Υπουργείου Παιδείας και Πολιτισμού.


Μιχάλης Οικονομίδης
Διευθυντής Λυκείου Πολεμιδιών



Appendix 3

Consent form given to teachers and students (In Greek)

ΔΗΛΩΣΗ ΣΥΓΚΑΤΑΘΕΣΗΣ ΓΙΑ ΣΥΜΜΕΤΟΧΗ

ΤΙΤΛΟΣ ΕΡΓΑΣΙΑΣ: Misfitting Students in Mathematics Tests

(Η δήλωση να συμπληρωθεί από τον/την συμμετέχοντα)

Παρακαλώ υπογραμμίστε
αυτό που ισχύει στην
περίπτωση σας

Έχετε ενημερωθεί για το σκοπό της έρευνας;

ΝΑΙ / ΟΧΙ

Σας δόθηκε η ευκαιρία να ζητήσετε διευκρινήσεις και να
συζητήσετε για την έρευνα;

ΝΑΙ / ΟΧΙ

Σας δόθηκαν ικανοποιητικές απαντήσεις σε όλες τις ερωτήσεις;

ΝΑΙ / ΟΧΙ

Με ποιόν μιλήσατε;

..... Π. Παπαφιδύ

Συγκατατίθεστε να λάβετε μέρος στην εργασία;

ΝΑΙ / ΟΧΙ

Συγκατατίθεστε να λάβετε μέρος σε συνεντεύξεις που θα
μαγνητοφωνηθούν;

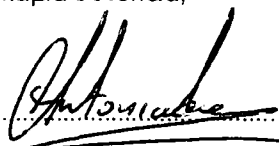
ΝΑΙ / ΟΧΙ

Αντιλαμβάνεστε ότι είσαστε ελεύθεροι να αποσυρθείτε από την εργασία:

- * οποιαδήποτε στιγμή το αποφασίσετε
- * χωρίς να χρειαστεί να δώσετε εξηγήσεις
- * χωρίς καμία συνέπεια;

ΝΑΙ / ΟΧΙ

Υπογραφή



Ημερομηνία

30/1/05.

(ΟΝΟΜΑΤΕΠΩΝΥΜΟ ΜΕ ΚΕΦΑΛΑΙΑ)

ΑΝΤΩΝΙΑΔΗΣ ΣΤΕΛΙΟΣ

Θέση: Καθηγητής / Μαθητής (Παρακαλώ υπογραμμίστε αυτό που ισχύει στην περίπτωση σας)

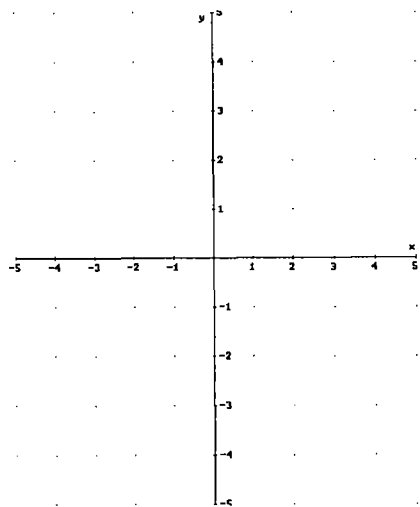
Appendix 4

Maths test in phase 1, on straight line graphs (Translated in English)

TEST ON STRAIGHT LINES

Directions: This test contains 12 questions. Answer all questions in the spaces provided.
If additional space is needed use the last blank page.

1. Plot the point **A(-2, 3)** on the axes provided.

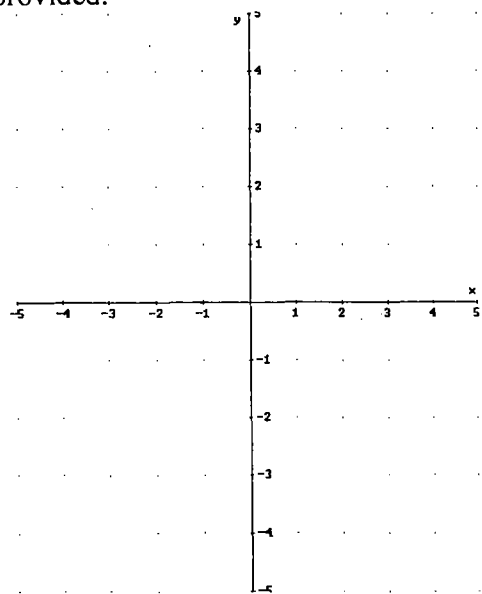


(2 m.)

2. Does the point **B(2, 3)** lies on the line $y = 2x + 1$? Show your working.

(4 m.)

3. Draw the line with equation $y = 2x - 3$ on the axes provided.



(4 m.)

4. Find the gradient of the line whose equation is $3y - 2x + 1 = 0$

(3 m.)

5. State the coordinates of the point where the lines with equations $x = 4$ and $y = 3$ meet.

6. Find the value of α for which the lines $y = (\alpha+1)x + 3$ and $y + 3x + 5 = 0$ are parallel. (2 m.)

7. Find the value of μ for which the lines $y = 2\mu x + 1$ και $y - \frac{1}{3}x = 2$ are perpendicular. (2 m.)

8. Find the equation of the line whose gradient is 3 and passes through the point (2,9). (3 m.)

(3 m.)

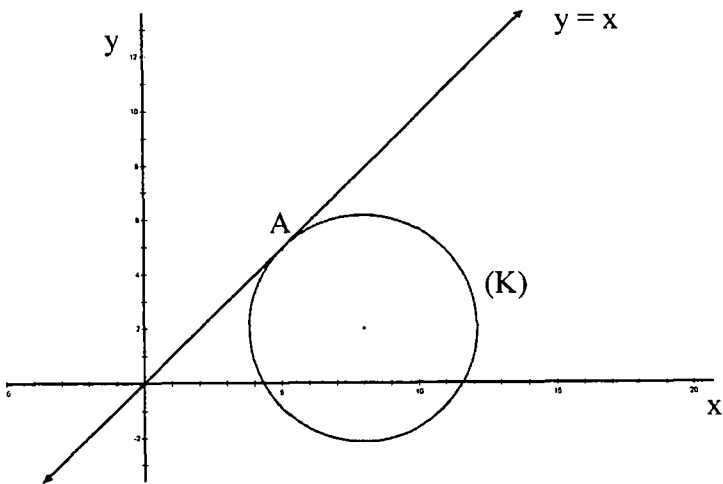
9. Find the equation of the line which passes through the point **(1, -2)** and is parallel to the x-axis.

(3 m.)

10. (α) Show that the equation of the line which passes through the points **(-1, 6)** and **(-2, 4)** is $y - 2x = 8$.
- (β) If this line crosses the x-axis at the point **A** and the y-axis at the point **B**, find the coordinates of the vertices of triangle **OAB**, where **O** is the origin.

(6 m.)

11. The diagram below shows the line with equation $y = x$, which is a tangent to circle (K) at point A.



Which of the following statements is correct? Circle the correct answer and hence give a reason for your answer.

- (α) The line $y = 0,95x$ cuts the circle at two points.
- (β) The line $y = 0,95x$ is also a tangent to the circle.
- (γ) The line $y = 0,95x$ does not have any common points with the circle.

Reason:.....

 (3 m.)

12. Find the equation of the line which is perpendicular to the line $3y - 9x + 1 = 0$ and cuts the y-axis at the same point as the line $y + 2(x+5) = 0$.

Appendix 5

Maths diagnostic test in phase 2 (Translated in English)

Maths Diagnostic test for 2005 - 06

Name:

Form:

ANSWER ALL THE QUESTIONS

1. Find the answers:

$$(\alpha) -3 + 12 =$$

$$(\beta) 4 \cdot (-3) =$$

$$(\gamma) (-12) \div (-6) =$$

$$(\delta) \frac{2}{3} + \frac{1}{5} =$$

$$(\epsilon) \left(-\frac{7}{3}\right) \cdot \left(\frac{6}{14}\right) =$$

(5 μ .)

2. Complete the following by placing $<$, $>$, or $=$ in the spaces provided

$$(\alpha) -3 \dots\dots\dots 2$$

$$(\beta) 2^3 \cdot 2^4 \dots\dots\dots 2^7$$

$$(\gamma) 3^5 \cdot 3^{-5} \dots\dots\dots \left(\frac{2}{3}\right)^0$$

(3 μ .)

3. Find and simplify the following:

$$(\alpha) x + x =$$

$$(\beta) x \cdot x =$$

$$(\gamma) (x^2)^3 =$$

$$(\delta) x^3 \div x^5 =$$

$$(\epsilon) (x-3)(x+2) =$$

$$(\sigma\tau) (2x-3)^2 =$$

$$(\zeta) x(3-2x) - (x-2)(x+2) =$$

(12 μ .)

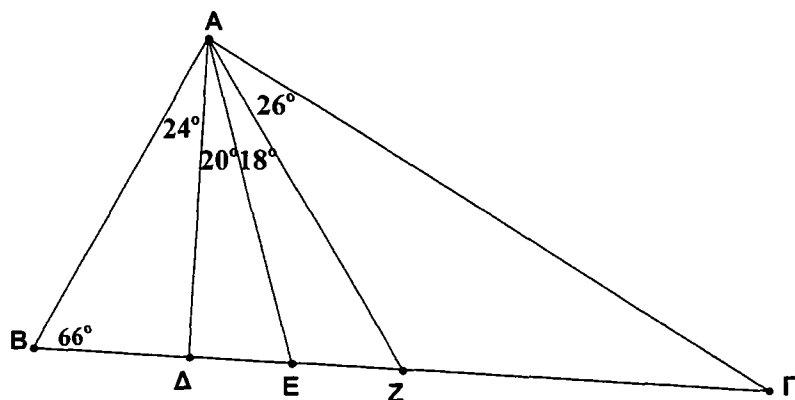
4. Simplify the following algebraic fractions:

$$(\alpha) \frac{2x}{x^2 + x} =$$

$$(\beta) \frac{x^2 - 7x + 12}{x^2 - 9} =$$

(5 μ .)

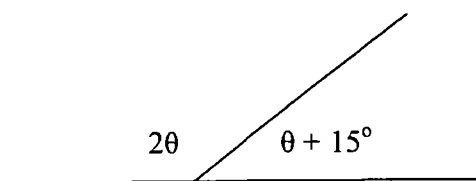
5. You are given triangle $AB\Gamma$ in which $BZ = Z\Gamma$. Complete the following:



(α) AZ is of $AB\Gamma$. (β) AE is of $AB\Gamma$.

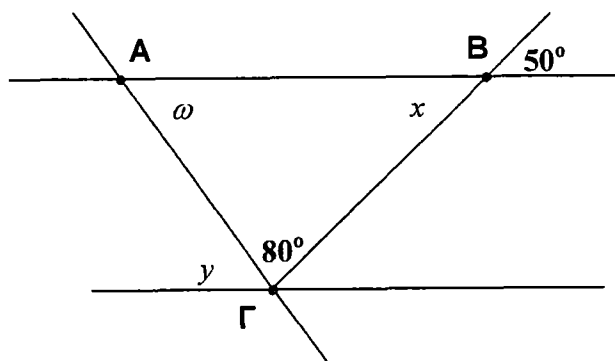
(γ) $A\Delta$ is of $AB\Gamma$. (δ) $\hat{\Gamma} = \dots\dots\dots$ (4 μ.)

6. With the help of the diagram given below find the value of y .



(2 μ.)

7. Calculate the angles labeled x, y και ω .



(α) $\hat{x} =$

(β) $\hat{y} =$

(γ) $\hat{\omega} =$

(δ) What kind of triangle is $AB\Gamma$;

Answer: (4 μ.)

8. Solve the equation: $\frac{x}{3} - 2 = 3x + \frac{1}{2}$

(3 μ.)

9. Given the equation $\frac{2x}{x-1} + \frac{3}{x+2} = 2$

(α) For which value(s) of x will the above equation have no roots; Απάντηση:

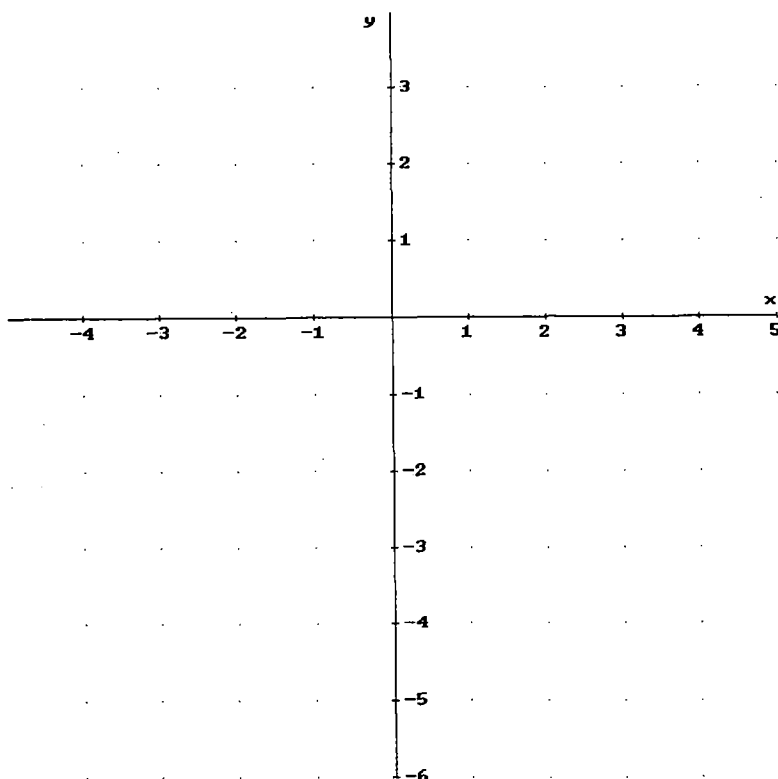
(β) Solve the equation.

(7 μ.)

10. If $y = 2x - 1$, complete the table below.
Hence sketch the graph of $y = 2x - 1$.

(5 μ.)

x	- 2	0	2
y			



Appendix 6

Maths test in phase 2, on quadratic equations (Translated in English)

Test on quadratic equations

PART A: Multiple choice questions (Questions 1 – 12)

Put a circle around the correct answer. Each question carries 1 mark.

1. The solutions of the equation $\alpha x^2 + \beta x + \gamma = 0$ are given by the formula:

(α) $\frac{-\beta \pm \sqrt{\beta^2 + 4\alpha\gamma}}{2\alpha}$

(β) $\frac{\beta \pm \sqrt{\beta^2 - 4\alpha\gamma}}{2\alpha}$

(γ) $\frac{-\beta \pm \sqrt{\beta^2 - 4\alpha\gamma}}{2}$

(δ) $\frac{-\beta \pm \sqrt{\beta^2 - 4\alpha\gamma}}{2\alpha}$

(ε) $-\beta \pm \frac{\sqrt{\beta^2 - 4\alpha\gamma}}{2\alpha}$

2. Given the equation $3x - 2x^2 - 5 = 0$, the values of α , β και γ respectively are:

(α) 2, 3, 5

(β) 3, -2, -5

(γ) -2, 5, 3

(δ) -2, 3, 5

(ε) -2, 3, -5

3. The sum of the roots of the equation $5x^2 - 3x + 2 = 0$ is:

(α) $\frac{5}{3}$

(β) $\frac{3}{5}$

(γ) $-\frac{3}{5}$

(δ) $\frac{2}{5}$

(ε) $-\frac{2}{5}$

4. The roots of the equation $2x^2 + 5x = 0$ are:

(α) 0 και -2,5

(β) 0 και 2,5

(γ) 0 και 0,4

(δ) 0 και -0,4

(ε) 2 και 5

5. The discriminant of the equation $2x^2 - 3x + 1 = 0$ is equal to:

(α) 0

(β) -1

(γ) 17

(δ) 1

(ε) 72

6. The roots of the equation $5x^2 - 10x + 4 = 0$

(α) Have a product of 4

(β) Have a sum of 10

(γ) Are real and equal

(δ) Are real and unequal

(ε) Are not real

7. Which of the following quadratic equations has roots $x_1 = 2$ and $x_2 = -7$?

(α) $2x^2 - 7x + 14 = 0$

(β) $x^2 - 5x - 14 = 0$

(γ) $x^2 + 5x - 14 = 0$

(δ) $x^2 - 5x + 14 = 0$

(ε) $x^2 + 5x + 14 = 0$

8. For which value of κ will the equation $x^2 + 3x - \kappa = 0$ have real and equal roots?

(α) 9

(β) 2,25

(γ) -9

(δ) $\frac{4}{9}$

(ε) -2,25

9. If $x = \frac{1}{2}$ is the one root of the equation $2x^2 + 3x - 2 = 0$, then the other one is:

- (α) 2 (β) -0,5 (γ) 3 (δ) -2 (ε) καμία από τις πιο πάνω

10. If the discriminant of $\alpha x^2 + \beta x + \gamma = 0$ is equal to 20, then the discriminant of $\gamma x^2 + \beta x + \alpha = 0$ is equal to:

- (α) 20 (β) -20 (γ) 0 (δ) $\frac{1}{20}$ (ε) cannot be calculated

11. If the sum of the roots of the equation $\kappa x^2 + 8x - 2\kappa = 0$ is equal to their product then the value of κ is:

- (α) $\kappa = -4$ (β) $\kappa = 2$ (γ) $\kappa = -2$ (δ) $\kappa = 8$ (ε) $\kappa = 4$

12. If $x_1 = 3$ and $P = 3$, then the value of S is:

- (α) -4 (β) -2 (γ) 0 (δ) 2 (ε) 4

PART B (Questions 13 – 16)

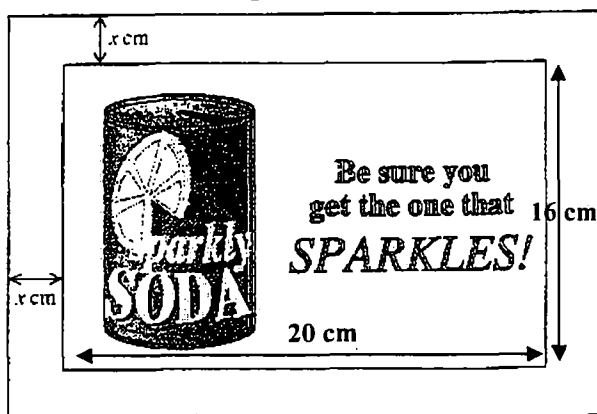
Answer the following questions in the spaces provided. Each question carries 4 marks.

1. Solve the equation $3x^2 + 5x - 2 = 0$.

2. For which value(s) of μ does the equation $2x^2 - 5 + 2x = \mu(x - 1)$ have roots that are opposite numbers?

3. If one root of the equation $x^2 - 2\mu x + 8 = 0$ is equal to the square of the other find the value(s) of μ .

4. figure 1



A rectangular advertising poster has dimensions 20cm and 16cm and is surrounded by a wooden frame of width x cm, as shown in figure 1.

(α) Show that the area of the wooden frame is given by the function

$$E(x) = 4x^2 + 72x$$

(β) Find the value of x for which the area of the wooden frame is 252 cm^2 .

Appendix 7

Test Anxiety Inventory (In Greek)

ΚΛΙΜΑΚΑ ΜΕΤΡΗΣΗΣ ΑΓΧΟΥΣ ΜΑΘΗΤΩΝ ΣΕ ΤΕΣΤ ΜΑΘΗΜΑΤΙΚΩΝ

ΟΔΗΓΙΕΣ

Πιο κάτω ακολουθούν κάποιες δηλώσεις τις οποίες χρησιμοποίησαν διάφοροι άνθρωποι για να περιγράψουν τους εαυτούς τους. Σημειώστε, ανάλογα με το πώς νιώθετε εσείς, σε ποια συχνότητα συμβαίνει σε εσάς αυτό που περιγράφει η κάθε δήλωση, βάζοντας κύκλο στον αντίστοιχο αριθμό που βρίσκεται δεξιά από κάθε δήλωση.

1 = Σχεδόν Ποτέ, 2 = Μερικές φορές, 3 = Συχνά, 4 = Σχεδόν Πάντα

Δεν υπάρχουν ορθές ή λανθασμένες απαντήσεις. Μην ξοδέψετε πολύ χρόνο σε μια δήλωση, αλλά δώστε την απάντηση η οποία νομίζετε ότι περιγράφει πώς εσείς νιώθετε.

Παρακαλώ απαντήστε όλες τις δηλώσεις.

Οι πιο κάτω δηλώσεις αναφέρονται σε διαγωνίσματα ή εξετάσεις στα ΜΑΘΗΜΑΤΙΚΑ

	1 ΠΟΛΥ ΛΙΓΗ	2 ΛΙΓΗ-ΠΟΛΥ ΣΥΧΝΗ	3 ΣΥΧΝΗ	4 ΠΟΛΥ ΣΥΧΝΗ
1. Αισθάνομαι αυτοπεποίθηση και ηρεμία κατά τη διάρκεια ενός διαγωνίσματος.....	1	2	3	4
2. Κατά τη διάρκεια εξετάσεων με διακατέχει ένα συναίσθημα ανησυχίας και αναστάτωσης.....	1	2	3	4
3. Το να σκέφτομαι το βαθμό ενός μαθήματος με επηρεάζει στο διαγώνισμα..	1	2	3	4
4. Σε σημαντικές εξετάσεις νιώθω να 'παγώνω'	1	2	3	4
5. Κατά τη διάρκεια των εξετάσεων συλλαμβάνω τον εαυτό μου να σκέφτεται αν ποτέ θα καταφέρω να τελειώσω το σχολείο.....	1	2	3	4
6. Όσο περισσότερο διαβάζω για ένα διαγώνισμα τόσο περισσότερο συγχύζομαι.....	1	2	3	4
7. Σκέψεις ότι δε θα πετύχω επηρεάζουν την αυτοσυγκέντρωσή μου στο διαγώνισμα	1	2	3	4
8. Αισθάνομαι να τρέμω όταν έχω ένα σημαντικό διαγώνισμα.....	1	2	3	4
9. Ακόμα και όταν είμαι καλά προετοιμασμένος/η για ένα διαγώνισμα νιώθω νευρικός/ή γι' αυτό.....	1	2	3	4
10. Νιώθω πολύ ανήσυχος/η μόλις πριν πάρω το διορθωμένο διαγώνισμά μου..	1	2	3	4

11. Κατά τη διάρκεια των διαγωνισμάτων νιώθω πολλή ένταση.....	1	2	3	4
12. Μακάρι οι εξετάσεις να μην με ενοχλούσαν τόσο πολύ.....	1	2	3	4
13. Κατά τη διάρκεια σημαντικών διαγωνισμάτων βρίσκομαι σε τόση ένταση που το στομάχι μου αναστατώνεται.....	1	2	3	4
14. Έχω την εντύπωση ότι στα σημαντικά διαγωνίσματα αποδίδω χειρότερα από όσα μπορώ.....	1	2	3	4
15. Αισθάνομαι πανικοβλημένος/η κατά τη διάρκεια ενός σημαντικού διαγωνίσματος.....	1	2	3	4
16. Ανησυχώ πάρα πολύ πριν από μια σημαντική εξέταση.....	1	2	3	4
17. Κατά τη διάρκεια των διαγωνισμάτων συλλαμβάνω τον εαυτό μου να σκέφτεται τις συνέπειες αποτυχίας.....	1	2	3	4
18. Κατά τη διάρκεια σημαντικών διαγωνισμάτων νιώθω την καρδιά μου να κτυπά πολύ γρήγορα.....	1	2	3	4
19. Αφού τελειώσει μια εξέταση προσπαθώ να μην ανησυχώ γι' αυτή αλλά δεν τα καταφέρνω.....	1	2	3	4
20. Κατά τη διάρκεια εξετάσεων είμαι τόσο νευρικός/η που ξεχνώ γεγονότα που σίγουρα ξέρω.....	1	2	3	4

ΜΕΡΟΣ Β: Άλλα στοιχεία

Παρακαλείστε να απαντήσετε τις πιο κάτω ερωτήσεις:

1. Τι βαθμό πήρατε στο πρώτο τρίμηνο στα Νέα Ελληνικά; Απάντηση:
2. Παρακολουθείτε απογευματινά ιδιαίτερα μαθήματα στα Μαθηματικά; (Ναι η Όχι)
Απάντηση:
3. Πόσο χρόνο αφιερώνετε περίπου κάθε μέρα για διάβασμα στα Μαθηματικά;
Δώστε την απάντησή σας σε λεπτά. Απάντηση:
4. Είναι τα Μαθηματικά ένα από τα αγαπημένα σας μαθήματα; (Ναι η Όχι) Απάντηση:

Appendix 8

Test Anxiety Inventory (Back translation from Greek)

Test Anxiety Inventory (translated from Greek)

Name Form

INSTRUCTIONS

Read below some of the statements that were used by several people to describe themselves under certain circumstances.

Please circle the relevant number on the right side of the page to show how you feel in similar situations.

1= almost never

2= some times

3= often

4= almost always

Please notice that there are no wrong answers. Do not spend time thinking over one statement. Give the best answer that describes your feelings.

Please circle all statements.

The following statements refer to tests in Maths.

1. I feel self-confident and relaxed during a test..... 1 2 3 4
2. During a test I feel anxious and upset..... 1 2 3 4
3. Thinking the marks is something that affects me in a test..... 1 2 3 4
4. In crucial tests, I freeze up..... 1 2 3 4
5. During the test I catch myself thinking if I will ever finish
school..... 1 2 3 4
6. The more I study for a test, the more confused I get..... 1 2 3 4
7. Thinking that I might fail affects my concentration on a test.... 1 2 3 4
8. Whenever I have a serious test, I tremble..... 1 2 3 4
9. Even when I am well prepared for a test I still feel nervous..... 1 2 3 4
10. I feel anxious just before I receive the corrected test..... 1 2 3 4
11. I feel very tensed during the tests..... 1 2 3 4

12. I wish I were not bothered so much by the tests..... 1 2 3 4
13. During important tests I am so nervous that I feel my
stomach upset..... 1 2 3 4
14. I think that on very important tests I perform less than I am
able to..... 1 2 3 4
15. I feel panicked on important tests..... 1 2 3 4
16. I feel very anxious before important tests..... 1 2 3 4
17. During a test I think about the consequences of failing..... 1 2 3 4
18. During important tests I feel my heart beating very fast..... 1 2 3 4
19. After the test I try not to worry but without success..... 1 2 3 4
20. I am so nervous during a test that I forget facts that I surely
know..... 1 2 3 4

Translated by Mr. Christos Constantinou

Mr. Constantinou is an assistant headmaster, head of the English department and an English teacher at the Lyceum of Polemidia.

Appendix 9

Maths Self- Esteem scale

.

Name: Form:

Mathematics Self-Concept Questionnaire

This questionnaire contains 6 statements that are more or less true (or false) descriptions of you. Please use the following six-point response scale to indicate how true (or false) each statement is as a description of you.

1 = Definitely False

2 = False

3 = More False than true

4 = More True than False

5 = True

6 = Definitely True

This is not a test. There are no right or wrong answers, and everybody will have different responses.

Please answer all items.

1. I am quite good at Mathematics.

2. I have generally done better in Mathematics courses than in other courses.

3. I have trouble understanding anything that is based on Mathematics

4. At school, my friends always come to me for help in Mathematics

5. I have never been excited about Mathematics.

6. I find many Mathematical problems interesting and challenging.

Thank you for your cooperation

Appendix 10

Maths Self- Esteem scale (In Greek)

Ονοματεπώνυμο:..... Τμήμα:

Ερωτηματολόγιο Αυτοεκτίμησης στα Μαθηματικά

Αυτό το ερωτηματολόγιο περιέχει έξι δηλώσεις οι οποίες είναι σε κάποιο βαθμό ορθές (ή λανθασμένες) περιγραφές του εαυτού σας. Παρακαλείστε να βάλετε στην γραμμή δεξιά από κάθε δήλωση τον αριθμό που αντιπροσωπεύει πόσο ορθά (ή λανθασμένα) περιγράφει η δήλωση εσάς.

1 = Απόλυτα Λανθασμένη

2 = Λανθασμένη

3 = Μάλλον Λανθασμένη

4 = Μάλλον Ορθή

5 = Ορθή

6 = Απόλυτα Ορθή

Αυτό είναι ερωτηματολόγιο και όχι διαγώνισμα. Δεν υπάρχουν ορθές ή λανθασμένες απαντήσεις και ο καθένας απαντά διαφορετικά. Μην ξοδέψετε πολύ χρόνο σε μια δήλωση, αλλά δώστε την απάντηση η οποία νομίζετε ότι ισχύει στην περίπτωση σας.

Παρακαλώ απαντήστε όλες τις δηλώσεις.

-
1. Είμαι αρκετά καλός στα Μαθηματικά. _____
 2. Γενικά τα πάω καλύτερα στα Μαθηματικά παρά σε άλλα μαθήματα. _____
 3. Έχω πρόβλημα στο να κατανοώ ότι βασίζεται στα Μαθηματικά. _____
 4. Στο σχολείο, οι φίλοι μου έρχονται πάντα σε μένα για βοήθεια στα Μαθηματικά. _____
 5. Ποτέ μου δεν ενθουσιάζομαι με τα Μαθηματικά. _____
 6. Βρίσκω κάποια Μαθηματικά προβλήματα ενδιαφέροντα και πρόκληση. _____

Σας ευχαριστούμε για τη συνεργασία σας

Appendix 11

ADHD scale (in Greek)

Κλίμακα Μέτρησης ADHD (Attention Deficit Hyperactivity Disorder)

Όνοματεπώνυμο μαθητή Τμήμα

Λύκειο

Να βάλετε σε κύκλο τον αριθμό που περιγράφει καλύτερα την συμπεριφορά του/της κάθε μαθητή/τριας.

0 = ΠΟΤΕ, 1 = ΜΕΡΙΚΕΣ ΦΟΡΕΣ, 2 = ΣΥΧΝΑ, 3 = ΠΟΛΥ ΣΥΧΝΑ

1. Κάνει απροσεξίες στις σχολικές εργασίες ή σε άλλες δραστηριότητες.....	0	1	2	3
2. Δυσκολεύεται να διατηρήσει την προσοχή του σε εργασίες στην τάξη ή σε άλλες δραστηριότητες.....	0	1	2	3
3. Δε φαίνεται να ακούει όταν απευθύνονται σε αυτόν.....	0	1	2	3
4. Δεν ακολουθεί μέχρι το τέλος οδηγίες και δεν διεκπεραιώνει εργασίες.....	0	1	2	3
5. Δυσκολεύεται στην οργάνωση εργασιών/καθηκόντων και δραστηριοτήτων.....	0	1	2	3
6. Αποφεύγει να εμπλακεί σε καθήκοντα/εργασίες που απαιτούν παρατεταμένη διανοητική προσπάθεια.....	0	1	2	3
7. Χάνει τα σχολικά του είδη (π.χ. πέννες, μολύβια, βιβλία).....	0	1	2	3
8. Αποσπάται η προσοχή του/της από εξωτερικούς παράγοντες.....	0	1	2	3
9. Ξεχνά καθημερινές δραστηριότητες.....	0	1	2	3
10. Κάνει νευρικές κινήσεις με τα χέρια ή τα πόδια ή κινείται στη θέση του/της.....	0	1	2	3
11. Κινείται συνεχώς στην τάξη ή σε άλλες περιπτώσεις στις οποίες αναμένεται να παραμένει στη θέση του/της.....	0	1	2	3
12. Είναι ανήσυχος/η.....	0	1	2	3
13. Δυσκολεύεται να εμπλακεί 'αθόρυβα' σε δραστηριότητες ή παιχνίδι.....	0	1	2	3
14. Είναι σε 'εγρήγορση' σαν να είναι 'συνεχώς αναμμένη η μηχανή του'.....	0	1	2	3
15. Μιλά περισσότερο απ' όσο χρειάζεται.....	0	1	2	3
16. Ξεστομίζει απαντήσεις πριν να ολοκληρωθούν οι ερωτήσεις.....	0	1	2	3
17. Δυσκολεύεται να περιμένει τη σειρά του/της.....	0	1	2	3
18. Διακόπτει ή επεμβαίνει όταν μιλούν άλλοι.....	0	1	2	3

Appendix 12

Questionnaire on Content Validity of the maths tests

Questionnaire on Content Validity of the Test

Having first considered the entire content of the unit on straight line graphs and the abilities and skills that should be possessed by students in the first form of the lyceum on this specific unit, please read the following statements and circle the number on the right of each statement which you feel represents the degree of your agreement or disagreement with each statement.

1 = Completely disagree, 2 = disagree, 3 = agree, 4 = absolutely agree

1. The format of the questions is appropriate for the students..... 1 2 3 4
2. All the questions are clear and unambiguous..... 1 2 3 4
3. Students who know the answers have enough time to finish the test... 1 2 3 4
4. All the important abilities and skills of the unit are assessed by the
test..... 1 2 3 4
5. No irrelevant topics are included in the test..... 1 2 3 4
6. The test content is representative of the unit content as described in
the curriculum..... 1 2 3 4
7. If your answer to any of the above questions is 1 or 2 please explain.
(e.g. which important skills of the unit are not assessed by the test?)

.....

.....

.....

.....

.....

.....

.....

Thank you for your help.

Panayiotis Panayides

Appendix 13

Questionnaire on Content Validity of the maths tests (In Greek)

ΕΡΩΤΗΜΑΤΟΛΟΓΙΟ

Αφού λάβετε πρώτα υπόψιν το κεφάλαιο στη γραφική παράσταση ευθείας και τις ικανότητες και δεξιότητες που πρέπει να κατέχει ένας μαθητής της Α΄ Λυκείου στο συγκεκριμένο κεφάλαιο, παρακαλώ διαβάστε τις πιο κάτω δηλώσεις και σημειώστε, βάζοντας σε κύκλο, το βαθμό που εσείς συμφωνείτε ή διαφωνείτε με την κάθε δήλωση.

1 = διαφωνώ απόλυτα, 2 = διαφωνώ, 3 = συμφωνώ, 4 = συμφωνώ απόλυτα

1. Η μορφή των ασκήσεων είναι κατάλληλη για τους μαθητές 1 2 3 4

2. Οι οδηγίες όλων των ασκήσεων είναι σαφείς 1 2 3 4

3. Οι μαθητές έχουν αρκετό χρόνο για να ολοκληρώσουν το διαγώνισμα 1 2 3 4

4. Το διαγώνισμα εξετάζει όλες τις σημαντικές ικανότητες και δεξιότητες
του κεφαλαίου 1 2 3 4

5. Το διαγώνισμα δεν περιλαμβάνει ασκήσεις από κεφάλαια άσχετα με το
συγκεκριμένο κεφάλαιο 1 2 3 4

6. Το περιεχόμενο του διαγωνίσματος είναι αντιπροσωπευτικό του
περιεχομένου του συγκεκριμένου κεφαλαίου, όπως περιγράφεται στο
αναλυτικό πρόγραμμα 1 2 3 4

7. Αν σε κάποια από τις πιο πάνω δηλώσεις σημειώσατε 1 ή 2, παρακαλώ δικαιολογείστε.
(π.χ. ποιες σημαντικές ικανότητες δεν εξετάζει το διαγώνισμα;)

.....

.....

.....

.....

.....

.....

Σας ευχαριστώ για τη βοήθεια σας

Παναγιώτης Παναγίδης

Appendix 14

Statistical Methods: Difference between two correlation coefficients and
log-linear analysis

Statistical methods

Fisher's Transformation

Given a correlation coefficient r , it can be transformed to r^* using Fisher's transformation:

$$r^* = \text{Fi}(r) \quad \text{where} \quad \text{Fi}(r) = 0.5 \cdot \ln\left(\frac{1+|r|}{1-|r|}\right) \quad \text{Also } \text{Fi}(-r) = -\text{Fi}(r)$$

Statement

Let r be the correlation coefficient of a bivariate random sample of size n , taken from a population having correlation coefficient ρ .

Then if $r^* = \text{Fi}(r)$ and $\rho^* = \text{Fi}(\rho)$: $r^* \sim N(\rho^*, \frac{1}{n-3})$ approximately, for large n (say $n \geq 50$).

Confidence Interval for the population correlation coefficient (ρ)

From the above, one can estimate a 95% confidence interval for the transformed correlation coefficient ρ^* , from r^* using:

$$\text{Lower limit } (\rho^*_L) = r^* - 1.96 \cdot \sqrt{\frac{1}{n-3}}, \quad \text{Upper Limit } (\rho^*_U) = r^* + 1.96 \cdot \sqrt{\frac{1}{n-3}}$$

From those limits, and using the inverse Fisher's transformation one can estimate a 95% confidence interval for the population correlation coefficient ρ .

Testing for differences between two correlation coefficients

Similarly, if r_1 and r_2 are the correlation coefficients of two bivariate samples of sizes n_1 and n_2 , taken from populations having correlation coefficients ρ_1 and ρ_2 , then:

$$r_1^* - r_2^* \sim N \left(\rho_1^* - \rho_2^*, \frac{1}{n_1 - 3} + \frac{1}{n_2 - 3} \right)$$

Hence the null hypothesis that $\rho_1^* = \rho_2^*$ can be tested using the test statistic:

$$Z = \frac{r_1^* - r_2^*}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

For large values of Z ($Z > 1.96$) the hypothesis is rejected and one can conclude that there is difference between ρ_1^* and ρ_2^* and therefore between ρ_1 and ρ_2 .

Log-linear analysis

The starting point for the analysis of nominal data on two or more attributes is a contingency table, each cell of which contains the frequency of occurrence of individuals in various combinations of categories.

When researchers are faced with crosstabulated data, their usual response is to compute a chi-square test of independence for each subtable. However, the chi-square test is insufficient when one has more than two categorical variables because it only tests the independence of the variables. It cannot detect various associations and interactions between the variables.

Log-linear analysis is a multivariate extension of the chi-square. It is a goodness-of-fit test that allows one to test all the effects (main effects, association effects and interaction effects) at the same time.

Log-linear models

Log-linear models are useful for uncovering the potentially complex relationships among the variables in a multiway crosstabulation. They are similar to multiple regression models.

Regression analysis examines the relationship between a dependent variable and a set of independent variables. Analysis of variance techniques provide tests for the effects of various factors on a dependent variable. But neither technique is appropriate for categorical data.

In log-linear models, all variables that are used for classification are independent variables and the dependent variable is the number of cases in a cell of the crosstabulation.

The basic idea of log-linear analysis is to search for the models that best fit the data. In order to do this, one needs to specify and compare all the models to each other. For this purpose, expected cell frequencies are generated for each model and the respective goodness-of-fit statistic is calculated.

Two chi-square statistics can be used:

The familiar Pearson chi-square statistic
$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

and the Likelihood-ratio chi-square
$$L^2 = 2 \sum_i \sum_j O_{ij} \cdot \ln \frac{O_{ij}}{E_{ij}}$$

For large sample sizes these statistics are equivalent. The advantage of the likelihood-ratio chi-square however, is that it can be subdivided into interpretable parts that add up to the total. This property is very useful when comparing the different models.

In the case of a contingency table with two variables A and B there are 5 possible models to be considered. If O_{ij} represents the observed frequency in the ij^{th} cell of the table, then:

Model 1: $\ln(O_{ij}) = \lambda$ where $\ln(O_{ij})$ is the natural logarithm of O_{ij} .

This model is one with no variable effect, with λ representing the overall mean effect.

Model 2: $\ln(O_{ij}) = \lambda + \lambda_{\alpha i}$.

This model represents the main effect of variable A.

Model 3: $\ln(O_{ij}) = \lambda + \lambda_{\beta j}$

This model represents the main effect of variable B.

Model 4: $\ln(O_{ij}) = \lambda + \lambda_{\alpha i} + \lambda_{\beta j}$

This model represents the main effect of variable A and the main effect of variable B and is called the independence model.

Model 5: $\ln(O_{ij}) = \lambda + \lambda_{\alpha i} + \lambda_{\beta j} + \lambda_{\alpha\beta}$

This model incorporates the overall mean effect, the main effects of both A and B and the interaction effect (association effect) of A and B.

First order interaction (involving two independent variables), in regression, occurs when an independent variable has different effects on a dependent variable at different levels of another independent variable. In other words, interaction means that the operation or influence of one independent variable on a dependent variable depends on the level of another independent variable. It is possible for three independent

variables to interact in their influence on a dependent variable. This is second order interaction. Higher order interactions are theoretically also possible but the higher the order the more difficult it is to interpret.

In log-linear analysis, where the dependent variable is simply the frequency in each cell, a first order interaction simply means association between the two independent categorical variables and second order interaction has the meaning of the association between two independent categorical variables depending on the different levels of another independent categorical variable.

The fifth and last model above is called the saturated model for the 2x2 contingency table.

In general, the saturated model derives its name by virtue of containing all the possible terms, including all main effects and all possible interactions between all variables. A goodness-of-fit test for the saturated model will always result in a chi-square value of zero because the saturated model possesses all the information among all the variables and thus will always perfectly reproduce the observed cell frequencies. This model is the basis for evaluating the goodness-of-fit of other models.

Assumptions

Log-linear analysis requires no distributional assumptions. The only assumption needed is that the observations are independent.

Furthermore, there are two requirements that are easy to satisfy in the present study:

- All the cells in the contingency table should have expected frequencies greater than 1
- No more than 20% of the cells can have expected frequencies less than 5.

The procedure

The different models need to be compared to determine which effects are associated with the differences in observed and expected frequencies. One has to always compare a model that contains a certain effect with one that does not, in order to appropriately determine which apparently has the greatest influence.

The most common procedure to approach the best model is called Backward Elimination. In this procedure one starts with the most complex model (usually the saturated model) and eliminates effects from it one by one in a step-wise fashion. The comparison between successive models is done by subtracting the L^2 value of one from the L^2 value of the other and the degrees of freedom of the one from the degrees of freedom of the other. Then critical values from the chi-square distribution can be used to evaluate the significance of the residual L^2 from the residual degrees of freedom.

Perhaps one confusing aspect of log-linear analysis is that a p-value above 0.05 (and not the usual $p < 0.05$ for significance) indicates that a given model fits the data adequately. This is because a p-value above 0.05 means lack of differences, indicating that the restricted model fits the data well and does not differ from the saturated model which contains all the variables being analyzed and all possible relationships between those variables (the restricted model contains only a subset of the possible relationships between the variables).

In such a case ($p > 0.05$) one will select the more parsimonious restricted model, because it can still 'explain' the data equally well as the saturated model while possess fewer relationships between the variables. In other words, the terms that were included in the saturated model but not in the restricted model can be dropped; they do not lend any explanatory value to the model.

Therefore the models are considered adequate if their significance level is at or above 0.05. However, best model is the one that accounts for the most effect in the data and at the same time is the most parsimonious.

Another way to approach the best model is to test for the significance of the individual terms in the model. A partial chi-squares table is produced by SPSS indicating the significance of each main effect, association or interaction term in the model. From that table one can choose all the significant terms to make the best and most parsimonious model.

Appendix 15

Log-linear analysis tables-results

Different schools * Different Teachers * Misfit

K-Way and Higher-Order Effects

	K	df	Likelihood Ratio		Pearson		Number of Iterations
			Chi-Square	Sig.	Chi-Square	Sig.	
K-way and Higher Order Effects(a)	1	129	1992,249	,000	3017,545	,000	0
	2	112	1770,638	,000	2424,666	,000	2
	3	48	,000	1,000	,000	1,000	2
K-way Effects(b)	1	17	221,611	,000	592,880	,000	0
	2	64	1770,638	,000	2424,666	,000	0
	3	48	,000	1,000	,000	1,000	0

a Tests that k-way and higher order effects are zero.

b Tests that k-way effects are zero.

Partial Associations

Effect	df	Partial Chi-Square	Sig.	Number of Iterations
school*teacher	48	1733,413	,000	2
school*testmisfit	4	,000	1,000	2
teacher*testmisfit	12	23,052	,027	2
school	4	100,698	,000	2
teacher	12	53,793	,000	2
testmisfit	1	67,121	,000	2

Interpretation: The first table shows significant 2-way effects. These are: school * teachers and teachers * misfit. There is significant association between teachers and misfit,

Teacher gender * Student gender * Misfit

K-Way and Higher-Order Effects

	K	df	Likelihood Ratio		Pearson		Number of Iterations
			Chi-Square	Sig.	Chi-Square	Sig.	
K-way and Higher Order Effects(a)	1	7	95,388	,000	94,881	,000	0
	2	4	4,455	,348	4,455	,348	2
	3	1	1,028	,311	1,028	,311	3
K-way Effects(b)	1	3	90,933	,000	90,427	,000	0
	2	3	3,427	,330	3,426	,330	0
	3	1	1,028	,311	1,028	,311	0

a Tests that k-way and higher order effects are zero.

b Tests that k-way effects are zero.

Interpretation: No evidence of a 2-way or 3-way effects.

Student gender * Ability * Anxiety * Misfit

K-Way and Higher-Order Effects

	K	df	Likelihood Ratio		Pearson		Number of Iterations
			Chi-Square	Sig.	Chi-Square	Sig.	
K-way and Higher Order Effects(a)	1	35	189,024	,000	181,370	,000	0
	2	29	102,647	,000	85,725	,000	2
	3	16	18,430	,299	18,565	,292	5
	4	4	,899	,925	,906	,924	4
K-way Effects(b)	1	6	86,377	,000	95,646	,000	0
	2	13	84,217	,000	67,160	,000	0
	3	12	17,531	,131	17,659	,126	0
	4	4	,899	,925	,906	,924	0

a Tests that k-way and higher order effects are zero.

b Tests that k-way effects are zero.

Partial Associations

Effect	df	Partial Chi-Square	Sig.	Number of Iterations
gender*Abilgroups30*Anxgroups30	4	3,427	,489	5
gender*Abilgroups30*testmisfit	2	3,037	,219	4
gender*Anxgroups30*testmisfit	2	5,066	,079	3
Abilgroups30*Anxgroups30*testmisfit	4	7,685	,104	4
gender*Abilgroups30	2	18,688	,000	3
gender*Anxgroups30	2	41,092	,000	4
Abilgroups30*Anxgroups30	4	38,420	,000	3
gender*testmisfit	1	,390	,532	5
Abilgroups30*testmisfit	2	4,639	,098	5
Anxgroups30*testmisfit	2	4,573	,102	5
gender	1	7,176	,007	2
Abilgroups30	2	15,897	,000	2
Anxgroups30	2	15,897	,000	2
testmisfit	1	47,407	,000	2

Interpretation: The first table shows significant 2-way effects. These are: gender * ability, gender * Anxiety and Ability * anxiety. No evidence of associations with misfit.

*Student gender * ADHD * Misfit*

K-Way and Higher-Order Effects

	K	df	Likelihood Ratio		Pearson		Number of Iterations
			Chi-Square	Sig.	Chi-Square	Sig.	
K-way and Higher Order Effects(a)	1	7	160,576	,000	171,082	,000	0
	2	4	21,576	,000	21,220	,000	2
	3	1	,019	,891	,019	,891	3
K-way Effects(b)	1	3	139,000	,000	149,861	,000	0
	2	3	21,557	,000	21,201	,000	0
	3	1	,019	,891	,019	,891	0

a Tests that k-way and higher order effects are zero.

b Tests that k-way effects are zero.

Partial Associations

Effect	df	Partial Chi-Square	Sig.	Number of Iterations
gender*adhd	1	20,271	,000	2
gender*testmisfit	1	1,349	,245	2
adhd*testmisfit	1	,478	,490	2
gender	1	1,200	,273	2
adhd	1	69,724	,000	2
testmisfit	1	68,077	,000	2

Interpretation: The first table shows significant 2-way effects. The only association evident is between gender and ADHD.

*Student gender * Study time * Misfit*

K-Way and Higher-Order Effects

	K	df	Likelihood Ratio		Pearson		Number of Iterations
			Chi-Square	Sig.	Chi-Square	Sig.	
K-way and Higher Order Effects(a)	1	11	563,047	,000	544,364	,000	0
	2	7	7,361	,392	7,223	,406	2
	3	2	4,977	,083	5,042	,080	3
K-way Effects(b)	1	4	555,686	,000	537,140	,000	0
	2	5	2,384	,794	2,181	,824	0
	3	2	4,977	,083	5,042	,080	0

a Tests that k-way and higher order effects are zero.

b Tests that k-way effects are zero.

Interpretation: The table shows no significant 2-way effects or 3-way effects.

*Student gender * Private tuition * Misfit*

K-Way and Higher-Order Effects

	K	df	Likelihood Ratio		Pearson		Number of Iterations
			Chi-Square	Sig.	Chi-Square	Sig.	
K-way and Higher Order Effects(a)	1	7	74,629	,000	81,124	,000	0
	2	4	8,125	,087	8,183	,085	2
	3	1	,222	,637	,222	,638	3
K-way Effects(b)	1	3	66,504	,000	72,941	,000	0
	2	3	7,903	,048	7,961	,047	0
	3	1	,222	,637	,222	,638	0

a Tests that k-way and higher order effects are zero.

b Tests that k-way effects are zero.

Partial Associations

Effect	df	Partial Chi-Square	Sig.	Number of Iterations
gender*pt	1	4,246	,039	2
gender*testmisfit	1	,016	,899	2
pt*testmisfit	1	3,667	,055	2
gender	1	6,945	,008	2
pt	1	12,699	,000	2
testmisfit	1	46,860	,000	2

Interpretation: The first table shows significant 2-way effects. The only association evident is between gender and private tuition.

*Student gender * Item order * Misfit*

K-Way and Higher-Order Effects

	K	df	Likelihood Ratio		Pearson		Number of Iterations
			Chi-Square	Sig.	Chi-Square	Sig.	
K-way and Higher Order Effects(a)	1	7	72,226	,000	71,748	,000	0
	2	4	2,468	,650	2,431	,657	2
	3	1	,003	,958	,003	,958	3
K-way Effects(b)	1	3	69,759	,000	69,317	,000	0
	2	3	2,465	,482	2,429	,488	0
	3	1	,003	,958	,003	,958	0

a Tests that k-way and higher order effects are zero.

b Tests that k-way effects are zero.

Interpretation: The table shows no significant 2-way effects or 3-way effects.

*Student gender * Atypical schooling * Misfit*

K-Way and Higher-Order Effects

	K	df	Likelihood Ratio		Pearson		Number of Iterations
			Chi-Square	Sig.	Chi-Square	Sig.	
K-way and Higher Order Effects(a)	1	7	339,257	,000	361,171	,000	0
	2	4	1,947	,745	1,933	,748	2
	3	1	,000	1,000	,000	1,000	2
K-way Effects(b)	1	3	337,309	,000	359,237	,000	0
	2	3	1,947	,583	1,933	,586	0
	3	1	,000	1,000	,000	1,000	0

a Tests that k-way and higher order effects are zero.
b Tests that k-way effects are zero.

Interpretation: The table shows no significant 2-way effects or 3-way effects.

*Student gender * Is maths favourite * Misfit*

K-Way and Higher-Order Effects

	K	df	Likelihood Ratio		Pearson		Number of Iterations
			Chi-Square	Sig.	Chi-Square	Sig.	
K-way and Higher Order Effects(a)	1	7	231,524	,000	177,500	,000	0
	2	4	3,444	,486	3,417	,491	2
	3	1	,000	1,000	,000	1,000	2
K-way Effects(b)	1	3	228,080	,000	174,083	,000	0
	2	3	3,444	,328	3,417	,332	0
	3	1	,000	1,000	,000	1,000	0

a Tests that k-way and higher order effects are zero.
b Tests that k-way effects are zero.

Interpretation: The table shows no significant 2-way effects or 3-way effects.

*Student gender * Revision periods before test * Misfit*

K-Way and Higher-Order Effects

	K	df	Likelihood Ratio		Pearson		Number of Iterations
			Chi-Square	Sig.	Chi-Square	Sig.	
K-way and Higher Order Effects(a)	1	11	297,668	,000	285,650	,000	0
	2	7	2,590	,920	2,506	,927	2
	3	2	,000	1,000	,000	1,000	2
K-way Effects(b)	1	4	295,078	,000	283,144	,000	0
	2	5	2,590	,763	2,506	,776	0
	3	2	,000	1,000	,000	1,000	0

a Tests that k-way and higher order effects are zero.

b Tests that k-way effects are zero.

Interpretation: The table shows no significant 2-way effects or 3-way effects.

